THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par : *l'Université Toulouse 3 Paul Sabatier (UT3 Paul Sabatier)*

Présentée et soutenue le *5 décembre 2016* **par :**

CHARLES COLAVOLPE

Étude des schémas de discrétisation temporelle « Explicite Horizontal, Implicite Vertical » dans une dynamique non-hydrostatique pleinement compressible en coordonnée masse

---

JURY

THOMAS DUBOS	Professeur chargé de cours	Président du Jury
PIERRE BÉNARD	Ingénieur en chef des Ponts, des Eaux et des Forêts	Membre du Jury
PIET TERMONIA	Directeur de recherche	Membre du Jury
LAURENT DEBREU	Chargé de recherche	Membre du Jury
SARAH-JANE LOCK	Chargée de recherche	Membre du Jury

École doctorale et spécialité :
SDU2E : Océan, Atmosphère et Surfaces Continentales

Unité de Recherche :
Centre National de Recherches Météorologiques - UMR 3589

Directeur(s) de Thèse :
Fabrice VOITUS et *Pierre BÉNARD*

Rapporteurs :
Thomas DUBOS, *Laurent DEBREU* et *Piet TERMONIA*

Résumé

La résolution numérique du système d'équations pleinement compressibles en vue de son utilisation pour des applications en Prévision Numérique du Temps (PNT) soulève de nombreuses questions. L'une d'elles porte sur le choix des schémas de discrétisation temporelle à mettre en œuvre afin de résoudre ce système de la manière la plus efficace possible, pour permettre la continuelle amélioration qualitative des prévisions. Jusqu'alors, les schémas de discrétisation temporelle basés sur des techniques semi-implicites (SI) étaient les plus couramment employés PNT, compte tenu de leur robustesse et de leur grande propriété de stabilité. Mais avec l'émergence des machines massivement parallèles à mémoire distribuée, l'efficacité de ces techniques est actuellement remise en question, car leur confortable plage de stabilité est obtenue au prix de l'inversion d'un problème elliptique tri-dimensionnel très gourmand en communications.

Ce travail de thèse vise à explorer d'autres méthodes de discrétisation temporelle, en remplacement des méthodes SI, s'appuyant sur des approches de type Horizontalement Explicite et Verticalement Implicite (HEVI). D'une part, ces approches s'affranchissent de la contrainte numérique imposée sur le pas de temps par la propagation verticale des ondes rapides supportées par le système, grâce au traitement implicite des processus verticaux. D'autre part, elles exploitent le paradigme de programmation voulant que chaque colonne verticale du modèle numérique soit traitée par un unique processeur. Ainsi, le traitement implicite de cette direction n'engendre aucune communication entre les processeurs. Cependant, bien que ces approches HEVI apparaissent comme une solution attractive, rien ne garantit que leur efficacité puissent être aussi compétitive que celles des schémas SI. Pour ce faire, ces schémas HEVI doivent permettre l'utilisation de pas de temps raisonnables pour une application en PNT.

L'objectif de ce travail de thèse est d'élaborer un schéma de discrétisation temporelle HEVI le plus efficace possible pour une utilisation en PNT, c'est-à-dire, un schéma qui autorise le plus grand pas de temps possible. Dans cette optique, deux voies ont été explorées : la première, issue des méthodes à pas de temps fractionné, a permis de revisiter et d'améliorer un schéma de discrétisation temporelle déjà proposé mais dont l'examen n'a jamais été approfondi dans la littérature ; il s'agit du schéma d'avance temporelle saute-mouton trapézoïdal. Il a été mis en évidence que l'ajout d'un simple filtre temporel d'usage commun en PNT, améliore grandement la stabilité de ce schéma, lui permettant ainsi, à moindre coût, de rivaliser en terme de stabilité avec le schéma Runge-Kutta explicite d'ordre 3. La seconde voie, plus récente, s'est avérée la plus prometteuse. Elle repose sur l'utilisation des méthodes Runge-Kutta Implicite-Explicite (RK-IMEX) HEVI. Au cours de l'étude, il a été tout d'abord mis en évidence certains problèmes de stabilité des schémas initialement suggérés dans la littérature en présence des processus d'advection. Puis, une nouvelle classe de schémas RK-IMEX HEVI s'appuyant sur un traitement temporel spécifique des termes d'ajustement horizontaux, a été proposée. Ce nouveau traitement remédie non seulement aux problèmes de stabilité ci-avant identifiés, sans surcoût numérique, mais permet également l'utilisation de plus grand pas de temps que ceux pouvant être envisagés via les méthodes HEVI à pas de temps fractionné. Outre le traitement des processus dynamiques affectant la propagation horizontale des ondes rapides, une étude annexe a mis en lumière l'apport bénéfique d'un traitement temporel implicite des termes non-linéaires orographiques (résultant de l'utilisation d'une coordonnée verticale épousant le relief) sur la stabilité du schéma HEVI. Enfin, tous les résultats théoriques obtenus ont été confrontés à des expérimentations numériques à l'aide d'un modèle plan vertical pour les équations pleinement compressibles en coordonnée masse.

Abstract

The use non-hydrostatic fully compressible modelling system in the perspective of Numerical Weather Prediction (NWP) raises many challenging questions, among which the choice time discretization scheme. It is commonly acknowledge that the ideal time marching algorithms to integrate the fully compressible system should both overcome the stability constraint imposed on time-step by the fast propagating waves supported by the system, and be scalable enough for efficiently computing on massively parallel computer machine. The assumed poor scalability property of Semi-implicit (SI) time schemes, currently favoured in NWP, is quite a drawback as they require global communications to solve a full three-dimensional elliptic problem. Because it is considered as the best compromise between stability, accuracy and scalability the properties of various classes of Horizontally Explicit Vertically Implicit (HEVI) schemes have been deeply explore in this work in a view of solving the fully system in mass-based coordinate. This class of time discretization approach eliminates all the problems linked to the implicit treatment of horizontal high-frequency forcings by coupling multi-step or multi-stage explicit methods for the horizontal propagation of fast waves to an implicit scheme for the treatment of vertically propagating elastic disturbances. The limitation in time-steps compared to SI schemes would be compensated by a much more economical algorithm per time-step. However, it is not firmly established that the efficiency of such a HEVI schemes could compete with one of the semi-implicit schemes.

The main objective of this Phd thesis work is to elaborate an efficient HEVI time scheme allowing usable time-step for NWP applications. For this purpose, the so-called explicit time-splitting technique and the recently suggested Runge-Kutta IMEX (RK-IMEX) schemes have been explored under HEVI approach. Firstly, the superiority in term of stability of the RK-IMEX methods in respect with the time-splitting approach has been confirmed. However, in presence of advection processes some unstable numerical behaviour of these schemes has been pointed out. To circumvent this problem a new class of RK-IMEX HEVI schemes has been proposed. This new class of HEVI time schemes reveals to be very attractive since they provide both good stability and accuracy properties. Secondly, in a side aspect of the HEVI approach, the stability impact of the temporal treatment of the terrain-following coordinate non-linear metric terms has been demonstrated. Numerical analyses on simplified framework indicate that there might be a benefit to deal with these specific terms in the implicit part of the HEVI schemes. All the theoretical studies have been confirmed by numerical testing through the use of a Cartesian vertical plane fully compressible model cast in a mass-based coordinate.

À Camille, ma fiancée,
son amour est une force.

À Isabelle & César, mes parents,
votre affection a porté ses fruits.

À mes grand-parents,
l'histoire que vous initiez, demeure.

À Raphaëlle & Damien et Pierre-Arthur & Vaianu, mes frères et sœurs,
vos expériences ont illuminé ma vie.

À Hélène & Jean-Michel, Alice & Glenn et Benjamin & Irina, ma belle-famille,
votre bienveillance me protège.

À Augustin & Étienne, mes neveux,
puissions-nous vous guider aussi bien.

À ma famille,

À mes amis,

pour votre affection et votre indéfectible soutien.

À notre directeur de thèse,
Le Docteur Fabrice VOITUS,
Ingénieur de Travaux en Météorologie,

vos idées lumineuses dans tous les domaines de votre travail ont initié de nombreux résultats importants. Votre soutien et votre compétence sont très présents dans ce travail.

Nous vous remercions et vous assurons de notre profonde reconnaissance.

À notre co-directeur de thèse,
Le Docteur Pierre BÉNARD,
Ingénieur en chef des Ponts, des Eaux et des Forêts,

vous nous avez guidé avec bienveillance. Votre sagesse et votre rigueur ont su être un facteur d'efficacité et de soutien.

Nous vous remercions et vous assurons de notre attachement.

À l'ensemble de l'équipe du GMAP,
À l'ensemble du personnel de Météo-France,

dont l'esprit si chaleureux est très caractéristique de Toulouse.

Table des matières

Table des figures

Liste des tableaux

Introduction

La *prévision numérique du temps* (PNT) a pour but de déterminer l'état de l'atmosphère pour une échéance fixée. Pour cela, elle cherche à approcher les solutions de systèmes d'équations décrivant la dynamique et la physique de l'atmosphère par l'utilisation de *schémas numériques*. Ces schémas visent à réaliser l'intégration temporelle des équations dynamiques par des manipulations algébriques de termes issus d'une discrétisation des opérateurs continus. Pour un système d'équations donné, le volume des calculs dépend, entre autre, du nombre d'états de l'atmosphère à calculer (dépendant de la durée d'intégration, du pas de temps et du schéma) ainsi que des techniques de discrétisations spatiales (notamment la taille du domaine et les résolutions dans les trois dimensions). Dans un contexte opérationnel, ces schémas sont soumis à des contraintes d'efficacité. Cette notion est complexe à définir car elle représente l'équilibre entre le volume des calculs du schéma et le temps d'exécution de l'algorithme. Ainsi, cette notion est fonction des schémas, mais aussi des ordinateurs sur lesquels ils sont appliqués. En effet, dans le cas où ces algorithmes sont appliqués par des machines massivement parallèles à mémoire distribuée utilisant plusieurs processeurs, le temps de communication entre ces processeurs nuit à l'efficacité de l'algorithme car elle augmente le temps d'exécution du programme. Cette contrainte informatique plaide en faveur de l'utilisation de méthodes locales, générant le minimum de communications. Par ailleurs, la montée en puissance offerte par l'utilisation des ces architectures permet d'envisager une forte augmentation des résolutions spatiales, ce qui, pour la PNT, reste une exigence continue. Une des issues pour maintenir une certaine efficacité des algorithmes consiste à avoir des pas de temps les plus grands possibles. Ceci permet d'obtenir moins d'états intermédiaires à calculer pour atteindre l'échéance fixée et ainsi tend à diminuer le volume de calcul à réaliser.

La durée maximale du pas de temps pour les schémas temporels dits *explicites*[1] ne peut être définie arbitrairement. Parmi les contraintes existantes, il en est une qui impose qu'une relation de proportionnalité existe entre le rapport du pas de temps et des résolutions spatiales de la discrétisation et l'inverse de la vitesse de propagation du processus le plus rapide supportés par les équations (contrainte dite de Courant-Fredrichs-Lewy *CFL* [12]). Si le pas de temps du schéma viole cette contrainte, alors la suite des états calculés diverge et aucune prévision n'est possible. Ainsi, pour ces schémas, le pas de temps est limité par deux paramètres : les résolutions spatiales et la vitesse des processus modélisés par le système d'équations.

Afin d'augmenter le pas de temps, plusieurs solutions ont été envisagées. L'une d'elles consiste à supprimer la présence des processus les plus rapides. Dans le cas d'un fluide compressible, tel que l'est l'atmosphère, ces *systèmes filtrés* visent à supprimer les ondes acoustiques (dues notam-

1. Schéma pour lequel l'état suivant se calcule uniquement en fonction des états déjà connus.

ment aux effets de compression et de détente du fluide) qui ont peu d'intérêts météorologiques. Parmi ces systèmes filtrés, le plus largement employé en PNT est le systèmes des équations primitives hydrostatiques qui s'avère très pertinent pour décrire les phénomènes d'échelles horizontales supérieures à la dizaine de kilomètres (pour lesquels l'hypothèse de l'écoulement hydrostatique reste valide). Toutefois, pour de plus fortes résolutions, Daley (1988) [14] montre que les effets non-hydrostatiques doivent être pris en compte. D'autres systèmes filtrés non-hydrostatiques ont été utilisés pour mieux représenter les phénomènes de plus petites échelles ; le système anélastique (Lipps & Hemler (1982) [41]) et le système pseudo-compressible (Duran (1989) [18]). Bien qu'étant utilisés couramment dans le domaine de la recherche météorologique (modèle Méso-NH de Lafore *et al.* (1997) [36]), ces systèmes présentent l'inconvénient de déformer certains modes d'importance météorologique majeure (notamment les modes de Rossby). De manière générale, tous les systèmes filtrant les ondes acoustiques proposés dans la littérature (*ex :* Dubos & Voitus (2014) [17], Arakawa & Konor (2009) [1]), hormis le système hydrostatique, imposent l'inversion d'un problème elliptique tridimensionnelle (de type équation de Poisson) non-local pour obtenir la valeur de la pression. De ce fait, la stratégie consistant à supprimer les ondes acoustiques du système se révèle discutable dans un contexte où le caractère local des algorithmes mis en jeux revêt une importance cruciale. En conséquence, ces faiblesses inhérentes aux systèmes filtrés favorisent l'utilisation du système d'Euler non-hydrostatique et pleinement compressible qui ne souffre d'aucune approximation et permet de représenter l'ensemble des phénomènes météorologiques, tant à l'échelle synoptique de l'ordre de la dizaine de kilomètres que les phénomènes d'échelle kilométrique voir même sub-kilométrique. La popularité de ce système ne cesse de croître dans le monde de la PNT, en témoigne les nombreux modèles à travers le monde : AROME (France, et le consortium ALADIN), ICON (Allemagne), NICAM (Japon) et WRF (États-Unis).

Système d'Euler pleinement compressible en coordonnée masse

Le système d'Euler pleinement compressible découle directement des lois fondamentales de la dynamique appliquées à un fluide atmosphérique assimilé à un gaz parfait. Ces lois se traduisent mathématiquement par l'équation de la quantité de mouvement (deuxième loi de Newton [2]), l'équation de continuité (équation d'Euler [3]), l'équation de la thermodynamique (premier principe thermodynamique), et l'équation d'état des gaz parfaits. En première approximation, l'atmosphère est généralement comparée à une pellicule mince enveloppant la Terre, et subissant les effets de l'accélération gravitationnelle g supposée constante. L'ellipticité de la Terre est négligée au détriment d'une forme sphérique. L'évolution dynamique d'un gaz parfait, non visqueux, soumis à un champ de gravité g dans un référentiel galiléen en rotation, est ainsi décrite sous sa forme vecteur-invariant par le système :

2. Isaac Newton (1643-1727) : philosophe, mathématicien, physicien, alchimiste, astronome et théologien anglais, puis britannique

3. Leonhard Euler (1707-1783) : mathématicien et physicien suisse

$$\rho\frac{d}{dt}\mathbf{V}+\nabla p=2\rho\Omega(\vec{e}_z\times\mathbf{V})+\mathbf{F}_{\mathrm{v}} \qquad \text{(mvt. horizontal)} \tag{1}$$

$$\rho\frac{d}{dt}w+\partial_z p+\rho g=\mathbf{F}_{\mathrm{w}} \qquad \text{(mvt. vertical)} \tag{2}$$

$$C_p\frac{d}{dt}T-\frac{1}{\rho}\frac{dp}{dt}=\mathcal{Q} \qquad (1^\circ \text{ loi thermo.}) \tag{3}$$

$$\frac{d}{dt}\rho+\rho\left(\nabla\cdot\mathbf{V}+\partial_z w\right)=0 \qquad \text{(continuité)} \tag{4}$$

$$p=\rho\,R\,T \qquad \text{(gaz parfaits)} \tag{5}$$

où $d/dt=\partial_t+\mathbf{V}\cdot\nabla+w\partial_z$ est la dérivée Lagrangienne[4], $\mathbf{V}$ représente le vecteur vent horizontal dont les composantes zonale et méridionale sont respectivement u et v, w est la vitesse verticale, ρ est la masse volumique (ou densité) de l'air, p la pression, T la température et ϕ le champ de géopotentiel tel que $\phi=gz$. l'équation pronostique de la pression s'obtient en combinant l'équation de continuité (4) avec l'équation de la thermodynamique (3) et l'équation d'état des gaz parfaits (5) :

$$\frac{1}{p}\frac{d}{dt}p+\frac{C_p}{C_v}\left(\nabla\cdot\mathbf{V}+\partial_z w\right)=\frac{\mathcal{Q}}{C_vT} \qquad \text{(pression)} \tag{6}$$

Par ailleurs, $\mathcal{Q}$ désigne le terme d'échange de chaleur, $\mathbf{F}_{\mathrm{v}}$ et $\mathbf{F}_{\mathrm{w}}$ représentent les forces de friction. Dans la mesure où nous nous intéresserons exclusivement aux seuls termes relatifs à la dynamique du système, ces contributions physiques (effets diabatiques ou visqueux) seront délibérément ignorées par la suite. Il en va de même des effets de l'eau dans l'atmosphère. Par conséquent la constante des gaz parfaits R, ainsi que les capacités calorifiques à volume et pression constants C_v et C_p seront ceux de l'air sec.

Ce système supporte l'ensemble des processus dynamiques influençant l'état de l'atmosphère météorologique : la propagation des ondes acoustiques, de l'onde de Lamb (mode externe acoustique horizontal), la propagation des ondes d'inertie-gravité et celle des ondes dites de Rossby associées aux effets de la rotation terrestre (autrement dit à la force de Coriolis[5], notée $\mathbf{F}_c=-2\rho\Omega(\vec{e}_z\times\mathbf{V})$, avec Ω est la vitesse angulaire de rotation et $\vec{e}_z$ le vecteur unitaire vertical), sans oublier le processus de transport par le vent. Là encore, pour les mêmes raisons qui nous ont poussées à négliger les termes $\mathbf{F}_{\mathrm{v}}$ et $\mathbf{F}_{\mathrm{w}}$, nous négligeons également $\mathbf{F}_c$. Les ondes du système sont principalement générées par les *termes d'ajustements* qui sont composés, d'une part, par le gradient de pression dans les équations (1)-(2), et d'autre part, par la divergence totale du vent dans les équations (4)-(6). Par ailleurs, le transport par le vent est modélisé par les termes advectifs présents dans la définition de la dérivée Lagrandienne.

4. Joseph Louis, comte de Lagrange (1736-1813) : mathématicien, mécanicien et astronome italien, naturalisé français

5. Gustave Gaspard de Coriolis (1792-1843) : mathématicien et ingénieur français

Il peut également s'écrire sous la forme flux pour permettre la conservation de certaines quantités. Cette forme implique de remplacer l'équation d'évolution de la température soit par l'équation pronostique de la température potentielle $\theta = T/\Pi$ (avec la fonction d'Exner $\Pi = (p/p_{00})^{R/C_p}$ et $p_{00} = 1000\,\text{hPa}$ une pression de référence) (Smolarkiewicz [63]), soit par l'équation pronostique de l'énergie interne (Satoh (2002) [56]). Par ailleurs, Klemp *et al.* (2007) [34] montrent que les quantités diagnostiques qui se définissent de manière non-linéaire par rapport aux variables pronostiques sont approchées par des valeurs ayant nécessairement des erreurs plus grandes que les variables pronostiques. Ils préconisent donc d'utiliser comme variables pronostiques celles que l'on veut détériorer le moins possible.

La plupart des modèles bâtis sur ces équations pleinement compressibles favorisent la coordonnée hauteur z, ou encore une coordonnée hybride, épousant le terrain, élaborée à partir de cette même coordonnée hauteur (Gal-Chen & Somerville (1975) [23]). Cependant, il existe des avantages à avoir une coordonnée verticale qui reflète la nature du milieu atmosphérique (Dutton (1986) [21]), notamment le fait que l'atmosphère terrestre soit proche de son équilibre hydrostatique plaide plutôt en faveur de l'utilisation de la pression hydrostatique comme coordonnée verticale. C'est d'ailleurs la coordonnée qui a été adoptée par certains modèles comme AROME ou WRF. Un de ses avantages, selon Phillips (1957) [52], est qu'elle assure une répartition approximativement équitable de la masse et de l'énergie entre les mailles.

En s'appuyant sur les travaux de Laprise (1992) [39], la quantité théorique π (souvent appelée *pression hydrostatique*) vérifiant la relation hydrostatique $\partial_z \pi = -\rho g$ est utilisée en tant que coordonnée verticale du système d'Euler pleinement compressible. Un des intérêts d'une telle coordonnée est qu'elle permet de passer facilement du système des équations non-hydrostatiques au système en équations primitives hydrostatiques (d'usage courant dans d'autres applications telle que le climat). Il est par ailleurs d'usage commun de construire à partir de cette coordonnée pression hydrostatique, une coordonnée dite *hybride* nommée η qui permet de lisser les effets du relief avec l'altitude. De manière générale, la définition de π est donnée par les fonctions A et B de η à valeurs dans $[0;1]$ telles que :

$$\pi(x, y, \eta, t) = A(\eta) p_{00} + B(\eta)\pi_s(x, y, t) \tag{7}$$

où π_s est la pression hydrostatique de la surface. Les fonctions A et B sont choisies de telle sorte qu'elles satisfassent $m = \partial_\eta \pi > 0$ où m définit la métrique verticale mesurant le passage de la coordonnée π à la coordonnée masse η. Ces fonctions doivent également vérifier les conditions au sommet $A(0) = B(0) = 0$ et à la surface $A(1) = 0$, $B(1) = 1$. Par la suite, l'indice s indiquera la valeur de la variable considérée à la surface, et inversement l'indice T désignera la valeur de la variable prise au sommet du modèle. L'utilisation d'une coordonnée suivant la topologie du terrain permet une prise en compte simple de l'effet du relief (appelé aussi *orographie*) dans les équations de la dynamique.

En appliquant les règles de transformation de coordonnée définies par Kasahara (1974) [32], le système d'Euler en coordonnée hauteur défini précédemment s'écrit en coordonnée masse de la manière suivante :

$$\frac{d}{dt}\mathbf{V} + RT\left(\frac{\nabla\pi}{\pi} + \nabla q\right) + \nabla\phi + \mathsf{Y} = 0 \tag{8}$$

$$\frac{d}{dt}w - \frac{g}{m}\partial_\eta\left[\pi(\mathrm{e}^q - 1)\right] = 0 \tag{9}$$

$$\frac{d}{dt}T + \frac{RT}{C_v}\left(\nabla\cdot\mathbf{V} - \frac{\mathrm{e}^q}{H}\frac{\pi}{m}\partial_\eta w + \mathrm{e}^q\mathsf{X}\right) = 0 \tag{10}$$

$$\frac{d}{dt}q + \frac{\dot\pi}{\pi} + \frac{C_p}{C_v}\left(\nabla\cdot\mathbf{V} - \frac{\mathrm{e}^q}{H}\frac{\pi}{m}\partial_\eta w + \mathrm{e}^q\mathsf{X}\right) = 0 \tag{11}$$

$$\partial_t\pi_s + \int_0^1 \nabla\cdot(m\mathbf{V})\mathrm{d}\eta = 0 \tag{12}$$

avec les relations diagnostiques :

$$\mathsf{X} = \frac{1}{RT}\frac{\pi}{m}\partial_\eta\mathbf{V}\cdot\nabla\phi \tag{13}$$

$$\mathsf{Y} = \frac{1}{m}\partial_\eta\left[\pi(\mathrm{e}^q - 1)\right]\nabla\phi \tag{14}$$

$$\frac{\dot\pi}{\pi} = \mathbf{V}\cdot\frac{\nabla\pi}{\pi} - \int_0^\eta \nabla\cdot(m\mathbf{V}) \tag{15}$$

$$m\dot\eta = B\int_0^1 \nabla\cdot(m\mathbf{V})\mathrm{d}\eta - \int_0^\eta \nabla\cdot(m\mathbf{V})\mathrm{d}\eta' \tag{16}$$

$$\phi = \phi_s - \int_1^\eta \frac{m}{\pi}\left(RT\mathrm{e}^{-q}\right)\mathrm{d}\eta' \tag{17}$$

où $d/dt = \partial_t + \mathbf{V}\cdot\nabla + \dot\eta\partial_\eta$ désigne la dérivée Lagrangienne, ∇ est le gradient horizontal à η constant, $\dot\eta$ la vitesse verticale covariante, $q = \ln(p/\pi)$ est une variable pronostique adimensionnée décrivant la déviation de la pression réelle p par rapport à π, et $H = RT/g$ une hauteur caractéristique.

Les conditions aux limites supérieure et inférieure du système sont données par : des conditions matérielles $\dot\eta_s = \dot\eta_T = 0$ (qui assurent la conservation de la masse), une condition élastique au sommet du modèle (qui se traduit par l'annulation de la déviation de pression non-hydrostatique au sommet $q_T = 0$ ou encore $p_T = \pi_T = \mathrm{Cte}$ [6]), et une condition rigide à la surface assurant que l'accélération verticale d'une parcelle d'air située à la surface $\dot w_s$ soit uniquement gouvernée par son mouvement horizontal $\mathbf{V}_s$ le long des surfaces rigides $z_s = \phi_s/g$. Cette dernière condition se traduit mathématiquement par la relation $\dot w_s = \mathbf{V}_s\cdot\nabla z_s$.

Les effets de l'orographie sur la dynamique apparaissent clairement au travers de la variabilité horizontale du relief, lequel est présent non seulement de manière explicite dans l'équation du mouvement horizontal (8), mais également dans les termes notés X et Y et qui apparaissent dans la divergence totale, et donc dans les équations de la température (10) et de la pression (11).

L'influence des conditions aux bords dans le système s'exprime directement au travers des évaluations intégrales. Il convient de noter le fait que la coordonnée masse fait naturellement intervenir des opérateurs intégraux sur la verticale dans l'évaluation de certains termes apparaissant

6. Dans la plupart des cas, la pression est supposée nulle au sommet

dans les équations pronostiques, notamment le terme d'auto-conversion hydrostatique $\dot{\pi}/\pi$, défini par la relation intégrale (15). Le géopotentiel (et donc la hauteur) se déduit de la connaissance du profil vertical de pression hydrostatique et de température par une intégration depuis la surface $\phi_s = gz_s(x, y)$ vers le niveau géopotentiel souhaité, selon la relation (17). La vitesse verticale covariante $\dot{\eta}$ est, elle aussi, déterminée via une relation intégrale décrite par (16). Par ailleurs, la coexistence d'opérateurs de dérivation et d'intégration verticale dans la dynamique du modèle constitue une des particularités de cette coordonnée. Or, pour une atmosphère donnée, la façon dont se propage physiquement les ondes rapides n'est en rien modifiée par le choix de la coordonnée verticale. L'équation de structure, celle qui régit la structure spatio-temporelle des ondes rapides supportées par le système, doit maintenir son caractère local et ceci même en coordonné masse. Par conséquent, ces opérateurs verticaux vérifient des propriétés remarquables de sorte que l'équation de structure soit indépendante de la coordonnée.

Comme pour le système en coordonnée hauteur, le système en coordonnée masse supporte donc également des ondes acoustiques. La résolution de ce système par des méthodes purement explicites impose au pas de temps de respecter la contrainte CFL portant sur la propagation dans toutes les direction de ces ondes. Par conséquent, il semble que les schémas qui puissent assurer une relative bonne efficacité de l'algorithme global soient ceux qui s'affranchissent de cette contrainte CFL. Dans un contexte où les architectures des ordinateurs ont pu permettre une telle approche, les schémas implicites se sont popularisés auprès de nombreux modèles de PNT et continuent toujours à être employés.

Problématique autour des schémas semi-implicites couramment employés en PNT

Une des propriétés des schémas implicites est de permettre l'utilisation de plus grands pas de temps par la suppression des conditions CFL portant sur la propagation des processus modélisés par les termes traités par ce type de schéma. L'idée sous-jacente des schémas semi-implicites (SI) est de s'affranchir de la contrainte CFL pesant sur la propagation des ondes rapides du système d'Euler par le traitement implicite des termes responsables de leur propagation dans toutes les directions. Ce traitement implique la résolution d'une équation elliptique de type Helmholtz [7] tridimensionnelle à chaque pas de temps du modèle.

En dépit des efforts développés afin d'élaborer des schémas temporels qui permettent l'utilisation de plus grands pas de temps, il faut garder à l'esprit que l'efficacité de ces algorithmes est aussi dépendante de l'architecture de la machine sur laquelle ils sont employés. En effet, bien que très stables, les schémas SI induisent des calculs non-locaux, dans le sens où ils nécessitent des informations sur une grande partie du domaine afin d'inverser le problème d'Helmholtz. Or, l'évolution des ordinateurs tend vers des architectures à mémoire de plus en plus distribuée, de sorte que le nombre de communications exigées risque fort d'altérer l'efficacité initialement atteinte par ces schémas. La scalabilité [8] des algorithmes devient une nouvelle contrainte d'efficacité qui s'applique sur les schémas à utiliser. Les récents travaux de Müller & Scheichl (2014) [48] montrent

7. Hermann Ludwig Ferdinand von Helmholtz (1821-1894) : physiologiste et physicien allemand

8. Capacité d'un algorithme à maintenir sa vitesse d'exécution par processeur lors d'une augmentation du nombre de tâches à effectuer

que la résolution de l'équation d'Helmholtz, générée par un schéma SI à coefficients homogènes sur l'horizontale, via des méthodes itératives multi-grilles (géométriques ou algébriques) souffre d'un problème de scalabilité faible. Cela signifie que l'algorithme perd en efficacité lors d'une augmentation simultanée du nombre de processeurs et du nombre de calculs à effectuer. En conséquence, à l'heure actuelle, il semble que la technique implicite dans toutes les directions risque de devenir de moins en moins efficace à long terme.

Par ailleurs, comme mentionné ci-dessus, il est d'usage courant de modéliser la propagation des ondes rapides par des opérateurs linéaires à coefficients homogènes sur l'horizontale (Robert *et al.*(1972) [55]). Or, l'état autour duquel est obtenu ce système linéaire repose sur des hypothèses très fortes, qui ne sont pas nécessairement vérifiées à chaque instant et à chaque endroit du domaine. De ce fait, il existe une différence importante entre le modèle non-linéaire complet et le système linéaire du schéma SI, de sorte que des résidus non-linéaires peuvent engendrer des instabilités pour lesquelles aucun pas de temps ne puisse être utilisé. Simmons *et al.* (1978) [59] ainsi que Bénard *et al.* (2003) [7] montrent que ces résidus peuvent être contrôlés par la variation du profil vertical de la température. L'introduction de ce paramètre a permis de protéger certains modèles de ces inconditionnelles instabilités. Toutefois, Bénard *et al.* (2005) [8] met en évidence le fait que la présence de certains termes liés à l'orographie impose également une inconditionnelle instabilité du schéma SI tel qu'il est mis en œuvre dans les modèles spectraux de méso-échelle. Or, du fait de l'augmentation croissante de la résolution horizontale, les pentes du modèle sont de plus en plus fortes (*ex :* résolution de 500 m pour la prévision du temps dans un environnement urbain). Ainsi, ces contraintes font peser un risque de plus en plus grand sur la viabilité des schémas SI à coefficients homogènes sur l'horizontale.

Au-delà des questions de scalabilité et de stabilité, l'utilisation des schémas SI se heurte à certains problèmes de précision. Ikawa (1988) [30], montre que le traitement implicite de l'horizontale tend à déformer les ondes de gravité qui revêtent une importance météorologique cruciale à meso-échelle. De plus, on peut s'interroger sur l'effet de l'utilisation de grands pas de temps pour modéliser des processus ayant une courte durée de vie (*ex :* ondes orographiques). En effet, il est possible que les erreurs cumulées par les schémas SI, qui ont nécessairement une mauvaise représentation de ces phénomènes de petite échelle temporelle, puissent dégrader sensiblement les prévisions numériques à moyen terme.

Vers des schémas d'intégration temporelle plus locaux : l'approche HEVI

Une des réponses à l'ensemble des faiblesses des schémas SI sus-mentionnées est, *a priori*, d'avoir des schémas d'intégration temporelle plus explicites. Cependant, certains processus dans la couche limite atmosphérique (*ex :* mélanges turbulents ou rayonnement) ont une importance capitale en PNT, car ils influent directement sur le temps sensible et nécessitent, par conséquent, une résolution de la verticale plus fine. De plus, le nombre actuel de niveaux verticaux, ainsi que la puissance de chaque processeur, permettent que chaque colonne ne soit traitée que par un seul processeur. Dans ce contexte, seules les directions horizontales sont gourmandes en communications. Ainsi, non seulement le traitement implicite de la verticale n'affecte en rien la scalabilité de l'algorithme, mais en plus, il libère le schéma de la contrainte CFL portant sur la propagation verticale des ondes

rapides (imposant un pas de temps bien plus court que celui défini par le traitement explicite de l'horizontale). Par conséquent, la recherche d'équilibre entre le plus grand pas de temps possible et la scalabilité de la méthode oriente vers des schémas *horizontalement explicites et verticalement implicites* (HEVI) en traitant de manière explicite (et donc locale) les directions nécessitant des communications et de manière implicite les termes responsables de la plus forte contrainte CFL sur le pas de temps.

Par rapport aux méthodes SI, l'introduction d'un traitement explicite de l'horizontale impose nécessairement des contraintes supplémentaires sur le pas de temps. Chaque processus issu d'un couplage entre des termes évalués explicitement ajoute une nouvelle contrainte sur la stabilité. Ainsi, pour un schéma HEVI, le but est de proposer un traitement particulier pour chaque terme responsable de la propagation horizontale de ces processus afin de pouvoir utiliser le pas de temps le plus grand possible. Une première réflexion porte sur le traitement des termes d'ajustement horizontaux responsables de la propagation horizontale des ondes rapides qui sont les ondes les plus contraignantes pour définir la stabilité du schéma HEVI. Les termes advectifs doivent également subir un traitement explicite, mais ceci doit, le moins possible, renforcer la contrainte déjà existante sur le pas de temps. Enfin, reste à proposer un traitement particulier des termes orographiques X et Y qui peuvent être considérés à la fois comme horizontaux et verticaux. En résumé, ces trois ensembles de termes, modélisant trois processus distincts, doivent subir un traitement spécifique pour maintenir le plus grand pas de temps possible.

Des schémas HEVI sont déjà en application dans certains modèles de PNT (notamment WRF, NICAM et ICON). Du fait du traitement explicite de l'horizontale, le pas de temps est notamment soumis à la contrainte CFL portant sur la propagation horizontale des modes externes, ce qui nécessite l'utilisation de pas de temps relativement petits. Dans le but de rendre ces schémas plus efficaces, la méthode du *pas de temps fractionné* (appelée en anglais *time-splitting*) vise à traiter avec deux pas de temps différents les processus rapides et les processus lents pour rendre les calculs plus économes (Klemp & Wihelmson (1978) [35]. Plus récemment, d'autres méthodes ont été élaborées, traitant tous les termes sur le même pas de temps et en se concentrant uniquement sur la distinction horizontale et verticale de la direction de propagation des processus. Ces schémas Implicite-Explicite (IMEX) HEVI, sont encore très peu utilisés de manière opérationnelle en PNT, car peu d'études sur leurs stabilités, lorsqu'ils sont appliqués au système d'Euler pleinement compressible, ont été réalisées.

Le caractère explicite du traitement de l'horizontale suggère qu'il existe toujours un pas de temps critique assurant la stabilité de ces schémas HEVI, contrairement aux schémas SI qui sont susceptibles d'être inconditionnellement instables en présence de fortes pentes. De plus, il est à espérer que l'utilisation de pas de temps plus modestes entraîne une meilleure représentation des phénomènes de petites échelles.

Objectifs de la thèse

Les schémas de type HEVI apparaissent comme une alternative séduisante dans le cas où les méthodes SI s'avéreraient inefficaces dans un contexte massivement parallèle où l'exigence de scalabilité imposerait une contrainte forte sur le choix du schéma d'intégration temporelle. Toutefois la stabilité de ces méthodes HEVI doit encore être éprouvée pour l'intégration du système d'Euler en coordonnée masse. Pour être utilisé de manière opérationnelle dans un modèle de PNT, le schéma HEVI candidat doit avant tout satisfaire deux exigences fondamentales :

- L'algorithme doit s'exécuter rapidement, afin de permettre la réalisation des prévisions. De plus, cette vitesse d'exécution doit pouvoir être maintenue, et ce malgré les évolutions continuelles de certains paramètres (comme la résolution) et la migration du modèle vers d'autres calculateurs.
- Le schéma doit garantir une certaine précision. Dans le cas de la PNT, il est souhaitable que cette précision soit au moins d'ordre 2 en temps. À l'heure actuelle, il est d'ailleurs suffisant d'avoir cet ordre de précision.

Le but de ce travail de thèse est d'élaborer un schéma d'intégration temporelle HEVI le plus efficace dans ce contexte opérationnel, ce qui, compte-tenu des exigences ci-dessus, revient à rechercher le schéma HEVI d'ordre 2 assurant l'utilisation du plus grand pas de temps possible. Dans ce travail, nous nous efforcerons d'étudier, dans un premier temps, la stabilité linéaire des différents schémas proposés dans la littérature. L'objectif est d'écarter les schémas HEVI jugés comme les moins stables (celles dont la contrainte CFL est la plus forte). Cette méthode d'analyse nous fournira un outil d'investigation permettant d'éventuelles améliorations de stabilité de certains schémas.

Dans le premier chapitre, sera présentée la méthode d'analyse de stabilité linéaire mise en œuvre dans tout ce travail de thèse. Les différents paramètres pertinents, qui seront utilisés afin de discriminer les schémas entre eux, seront définis. Au terme de ce chapitre, le lecteur sera capable d'appréhender les différentes méthodes numériques mises en place dans les chapitres suivants.

Le deuxième chapitre présentera les études de stabilité sur la première famille de schémas HEVI mis en place dans la PNT : la méthode au pas de temps fractionné. Très populaires car faciles à mettre en œuvre, ces schémas sont bâtis de manière à être le plus efficace possible en économisant au maximum le calcul de certains termes, supposés évoluer lentement. Dans la littérature, il existe très peu d'articles énumérant et comparant la stabilité et la précision des schémas entre eux. Par ailleurs, ces études n'ont été menées que dans le cas de systèmes linéaires encore plus simplifiés que celui du système linéaire pleinement compressible linéarisé. Ainsi, beaucoup de schémas sont donc encore à étudier, particulièrement pour le système d'Euler en coordonnée masse dont la présentation n'a été faite que par Klemp (2007) [34], pour lequel il ne teste qu'un seul schéma. De plus, nous présentons dans ce chapitre une approche originale de cette méthode qui nous a permis d'améliorer la stabilité sans alourdir le schéma de calculs supplémentaires. Enfin, nous y proposons un nouveau schéma qui, d'après nos études, est plus efficace que ceux actuellement utilisés.

Dans le troisième chapitre, nous analyserons la seconde famille des méthodes HEVI : les schémas IMEX. Bien que cette famille très large de schémas existe depuis l'introduction des méthodes SI, leur étude pour le système complet d'Euler sous la contrainte HEVI est beaucoup plus récente. En effet, il n'existe que très peu de modèles opérationnels utilisant ce type de schéma, et aucun en

coordonnée masse. Très récemment, des études comparant de nombreux schémas IMEX HEVI ont été réalisées pour le système des ondes acoustiques 2D. Dans ce chapitre, nous allons reprendre ces études pour les étendre au système d'Euler pleinement compressible en coordonnée masse linéarisé. Nous allons voir que les perspectives issues des conclusions des précédentes études se heurtent à des instabilités issues des processus advectifs. Une étude de ce comportement permet d'introduire une nouvelle classe de schémas que nous démontrerons comme étant précises, plus stables, et sans nécessiter plus de calculs.

La quatrième partie se concentrera sur l'application de ces méthodes pour le système d'Euler complet. Cela soulèvera la question du traitement des termes non-linéaires liés à la présence d'orographie, qui sont totalement absents du système linéaire utilisé précédemment. Leur étude spécifique est notamment motivée par le fait que ces termes rendent le schéma très instable. Pour comprendre ce comportement, et ainsi être capable de proposer une solution pour améliorer la stabilité des schémas HEVI, des études théoriques sur un système simple et sur le système linéaire discrétisé nous guiderons vers une solution pour renforcer la stabilité de ces schémas en présence d'orographie. De plus, nous y présenterons les techniques de résolution du problème inverse imposé par le traitement implicite de la verticale.

Le dernier chapitre se concentrera sur la validation de l'ensemble de ces résultats théoriques par des expériences numériques réalisées par un modèle plan vertical développé pour mesurer les différences entres les meilleurs schémas retenus, ainsi que les différents traitements proposés tout au long de ce travail. Nous y présenterons la confirmation de nos études théoriques. Fort de cette vérification expérimentale, nous pouvons résumer l'ensemble des travaux de thèse dans une conclusion qui fera non seulement la synthèse de tous ces travaux, mais également les perspectives d'amélioration et de développement dans un vrai modèle de PNT opérationnel.

Chapitre I

Méthodologie d'analyse de stabilité linéaire d'un système en coordonnée masse

En l'absence de forçages externes, un schéma de discrétisation temporelle est dit stable, au sens de Hadamard [1], si « *la densité d'énergie de la solution discrète reste bornée au cours de l'intégration numérique* ». Autrement dit, l'amplitude de la solution numérique ne doit pas croître de manière exponentielle au fil de l'intégration, rendant ainsi toutes tentatives de prévisions impossibles. Toutefois, cette condition ne suffit pas à garantir la qualité de la prévision, il convient également que la solution ne soit pas amortie de façon dommageable et que sa phase ne soit pas altérée (ni par accélération ni par ralentissement) de façon exagérée. Ces *erreurs de phase*, traduisant la capacité du schéma numérique à représenter fidèlement les différentes caractéristiques en terme de propagation de la solution physique, peuvent s'accumuler durant la période d'intégration et devenir significatives sur de très longues périodes. Dans certains cas, ces erreurs peuvent être plus redoutables encore. En effet, si l'onde modélisée par le schéma possède une erreur de phase telle que sa vitesse de groupe s'annule, alors il y a un risque d'accumulation d'énergie en un point du modèle qui peut engendrer des instabilités numériques.

Afin d'analyser les propriétés de stabilité et de précision d'un schéma temporel, la méthode d'analyse linéaire de Von Neumann [2] a fait ses preuves. Cette dernière est fondée sur l'idée que la solution d'un système dynamique linéaire sur un domaine non-borné peut se décomposer comme une somme de modes propres ondulatoires (de Fourier). Ainsi, pour étudier les propriétés d'une discrétisation temporelle appliquée à ce système, il suffit d'examiner la stabilité et l'erreur de phase associées à chacun de ces modes propres ondulatoires pris individuellement.

L'objectif poursuivi dans ce chapitre est donc de décrire le principe de l'analyse appliquée au cas du système d'Euler en coordonnée masse, pour lequel l'influence des bords supérieur et inférieur est explicitement intégrée à la dynamique via la présence d'opérateurs intégraux sur la verticale. Les outils permettant de mieux appréhender les différentes techniques d'analyses numériques mises en œuvre dans ce travail de recherche seront ainsi fournies au lecteur.

1. Jacques Salomon Hadamard (1865-1963) : mathématicien français
2. John von Neumann (1903-1957) : mathématicien et physicien américano-hongrois

1 Principe de l'analyse continue en espace et discrète en temps

Considérons un système dynamique décrivant l'évolution temporelle de l'atmosphère. Il peut s'énoncer symboliquement sous la forme :

$$\partial_t X = \mathscr{M}(X)$$

où $X = X(x, y, \eta, t) = X(r, t)$ est le vecteur d'état, qui est constitué des P variables pronostiques $X = (X_1, \ldots, X_P)$ du système, et $\mathscr{M}$ représente l'opérateur différentiel non-linéaire contenant l'ensemble des processus dynamiques supportés par le système.

Le principe de l'analyse de Von Neumann suppose donc, dans un premier temps, la linéarisation du problème complet décrit par l'opérateur $\mathscr{M}$. La procédure de linéarisation classique dans le cadre d'un système dynamique atmosphérique consiste à choisir un état de référence stationnaire $\overline{X}$ (*ie :* $\partial_t \overline{X} = \mathscr{M}(\overline{X}) = 0$) et homogène dans l'espace, auquel on superpose un état perturbé $X' = X - \overline{X}$. Ainsi, l'opérateur linéaire associé au système est tel que :

$$\overline{\mathscr{L}} \cdot X = \mathscr{M}(\overline{X}) + \left.\frac{\partial \mathscr{M}}{\partial X}\right|_{\overline{X}} \cdot (X - \overline{X}) = \left.\frac{\partial \mathscr{M}}{\partial X}\right|_{\overline{X}} \cdot X' \tag{I.1}$$

Par la suite, nous omettrons le caractère primé de X' pour X en sachant pertinemment que nous parlerons de perturbation autour de l'état de référence lorsque nous désignerons X lors de l'analyse. De même, par commodité d'écriture, nous désignerons $\overline{\mathscr{L}}$ simplement par $\mathscr{L}$ tout en gardant à l'esprit que cet opérateur linéaire est à coefficients constants dépendant de l'état de référence $\overline{X}$. Le problème linéarisé se met symboliquement sous la forme :

$$\partial_t X = \mathscr{L} \cdot X \tag{I.2}$$

À ce stade de la méthode, le caractère non-borné du domaine est crucial car il garantit l'admissibilité de modes oscillatoires pour le système linéarisé. Cependant, comme nous l'avons fait remarquer dans l'introduction, l'une des particularités du système en coordonnée masse réside dans le fait que les bords inférieur et supérieur influent directement sur la dynamique, au travers des évaluations diagnostiques faisant appel à des intégrales sur la verticale.

Afin de s'affranchir de ces conditions aux bords et ainsi permettre l'analyse à la manière de Von Neumann, la démarche adoptée est celle proposée par Bénard (2002) [6]. Elle consiste à les éliminer en appliquant à $\mathscr{L}$ les opérateurs différentiels verticaux l appropriés. Le système ainsi transformé se réécrit :

$$\partial_t (lX) = l\mathscr{L} \cdot X \tag{I.3}$$

où l est la matrice de diagonale constituée des $(l_1, \ldots, l_P)$ opérateurs spatiaux linéaires, tels que $(l_1\mathscr{L}_{11}, \ldots, l_P\mathscr{L}_{PP})$ deviennent des opérateurs linéaires qui ne contiennent plus de références aux

bords. Finalement, le système linéaire $l\mathscr{L}$ écrit sous une forme matricielle

$$\partial_t \begin{pmatrix} l_1 X_1 \\ \vdots \\ l_P X_P \end{pmatrix} = \begin{pmatrix} l_1\mathscr{L}_{11} & \cdots & l_1\mathscr{L}_{1P} \\ \vdots & \ddots & \vdots \\ l_P\mathscr{L}_{11} & \cdots & l_P\mathscr{L}_{1P} \end{pmatrix} \cdot \begin{pmatrix} X_1 \\ \vdots \\ X_P \end{pmatrix}$$

devient un système non-borné.

De manière évidente, l'une des premières conditions à remplir est la suivante :

[C1] : *il existe des opérateurs linéaires l tels que le système linéaire $l\mathscr{L}$ soit non-borné.*

À supposer que de tels opérateurs l existent, et donc que la condition [C1] soit satisfaite, trois autres conditions doivent être respectées avant de pouvoir appliquer la méthode d'analyse de Von Neumann au système non-borné $l\mathscr{L}$. Ces conditions sont énoncées ci-dessous :

Conservation de l'énergie :
Pour toute perturbation $X(t=0)$ autour de $\overline{X}$ avec une densité d'énergie bornée, l'évolution en temps $X(t)$ selon (I.3) doit avoir une densité d'énergie bornée.
Soit $X(r)$ un mode normal complexe du système linéaire non-borné $l\mathscr{L}$, vérifiant $l\mathscr{L} \cdot X(r) = \lambda l X(r)$ où λ est un complexe. L'évolution de ce mode est bornée en temps si, et seulement si, $\lambda \in \hat{\imath}\mathbb{R}$ [3]. Par conséquent, la condition de conservation de l'énergie du système non-borné peut être énoncée de la manière suivante :

[C2] : *Pour tout mode propre complexe de $l\mathscr{L}$, $\lambda \in \hat{\imath}\mathbb{R} \iff X(r)$ a une densité d'énergie bornée.*

Équivalence des systèmes borné et non-borné au sens des modes normaux :
Tout mode normal du système borné d'origine $\mathscr{L}$ est aussi mode normal du système non borné $l\mathscr{L}$, avec la même fréquence notée ω.
Soit $X(r) = \hat{X} f(r)$ un mode normal continu en temps du système borné vérifiant $\mathscr{L} \cdot X(r) = \hat{\imath}\omega X(r)$, où $\hat{X} = (X_1, \ldots, X_P) \in \mathbb{C}^P$ et $f = (f_1, \ldots, f_P) \in \mathbb{C}^P$, où f représente la structure et $\hat{X}$ la polarisation du mode. Alors, ce même mode normal vérifie $l\mathscr{L} \cdot X(r) = -\hat{\imath}\omega l X(r)$ pour le système non-borné. Autrement dit :

[C3] : *pour tout mode normal du système linéaire non borné de structure $f(r)$, $l_i f_i(r)$ doit être proportionnel à $f_i(r)$: $\forall i \in [\![1; P]\!]$, $\quad l_i f_i(r) = \xi_i f_i(r)$, avec $\xi_i \in \mathbb{C}$.*

[C4] : *pour tout mode normal du système linéaire non borné de structure $f(r)$, $l_i\mathscr{L}_{ij} f_j(r)$ doit être proportionnel à $f_i(r)$: $\forall (i,j) \in [\![1; P]\!] \times [\![1, P]\!]$, $\quad l_i\mathscr{L}_{ij} f_j(r) = \mu_{ij} f_j(r)$, avec $\mu_{ij} \in \mathbb{C}$.*

Ces deux dernières conditions permettent d'écrire chaque équation pronostique discrétisée en temps comme une équation scalaire pour tout mode normal du système linéaire non borné.

3. $\hat{\imath}$ est l'unité imaginaire telle que $\hat{\imath}^2 = -1$

Finalement, si les conditions [C1]-[C4] sont satisfaites, alors la stabilité linéaire du système discret borné se ramène (effets des bords mis à part) à l'analyse du système discret non-borné. En considérant les modes propres ondulatoires du système non-borné de la forme $X = \hat{X}(t)f(r)$, tels que pour chacun de ces modes, le système $l\mathscr{L}$ se réduit à une équation d'oscillation de la forme :

$$\partial_t \hat{X} = l\mathscr{L} \cdot \hat{X} = -\hat{\imath}\omega\hat{X}$$

La solution analytique est donnée par $\hat{X}_{\text{exact}}(t) = \hat{X}_{\text{exact}}(0)e^{\hat{\imath}\omega t}$, de sorte qu'elle vérifie la relation $\hat{X}_{\text{exact}}(t + \Delta t) = e^{-\hat{\imath}\omega\Delta t}\hat{X}_{\text{exact}}(t) \equiv \lambda_{\text{exact}}\hat{X}_{\text{exact}}(t)$ où $\lambda_{\text{exact}} = e^{-\hat{\imath}\omega\Delta t}$ et désigne le facteur d'amplification analytique du système au bout d'un pas de temps Δt qui, dans le cas de l'équation d'oscillation, est un complexe de module 1 ($|\lambda_{\text{exact}}| = 1$) et d'argument $-\omega$ (correspondant à la fréquence de l'onde). De plus, au bout de Δt, la phase du mode analytique est de $\theta_{\text{exact}} = \omega\Delta t$ (en radians).

Stabilité

L'analyse de Von Neumann examine la stabilité du système discrétisé en temps pour des perturbations qui ont la structure des modes propres continus en espace définis par $f(r)$. Pour ce faire, on émet l'hypothèse que l'amplitude des modes propres du système $X = \hat{X}(t)f(r)$ croit linéairement d'un pas de temps à l'autre d'un facteur d'amplification complexe λ, tel que :

$$\hat{X}^+ f(r) = \lambda\, \hat{X}^0 f(r) \tag{I.4}$$

$$\hat{X}^0 f(r) = \lambda\, \hat{X}^- f(r) \tag{I.5}$$

où $\hat{X}^-$, $\hat{X}^0$, et $\hat{X}^+$ sont respectivement les valeurs du vecteur d'état aux différents pas de temps $t - \Delta t$, t, et $t + \Delta t$ pour $t \geq 0$. Le schéma est dit *stable* si le module du taux de croissance λ est inférieur à 1, (*ie :* si $|\lambda| \leq 1$ pour tous les modes propres du système). Dans le cas où $|\lambda| = 1$, le schéma est dit neutre ; si $|\lambda| < 1$ le schéma est dit *amortissant*. Le schéma est dit *instable* si $|\lambda| > 1$.

Erreur de phase

L'hypothèse de croissance linéaire des modes propres du système, permet de relier directement la phase de la solution numérique noté θ à l'argument de son taux de croissance λ via la relation

$$\theta = \arctan\left(\text{Im}\left\{\lambda\right\} / \text{Re}\left\{\lambda\right\}\right) = -\arctan\left(\hat{\imath}\,\frac{\lambda - \lambda^*}{\lambda + \lambda^*}\right) \tag{I.6}$$

où λ^* désigne le complexe conjugué de λ. Soit $R = \theta/\theta_{\text{exact}}$ le rapport de l'avance de phase produite par le schéma numérique au bout d'un pas de temps par le changement de phase de la solution analytique au cours de ce même intervalle de temps. Si $R < 1$, le schéma est dit *ralentissant* ; si $R > 1$, le schéma est dit *accélérant*.

2 Application au système d'Euler en coordonnée masse

L'introduction rappelle que les équations d'Euler pleinement compressibles en coordonnée masse sont définies à partir de la coordonnée généralisée η. Toutefois, par simplicité, nous utiliserons pour toutes nos analyses la coordonnée masse « pression hydrostatique normalisée », ou encore appelée σ, définie par $\sigma = \pi/\pi_s$. Cette coordonnée est un cas particulier de la coordonnée η, qui elle possède une métrique plus simple, et facilite ainsi les analyses théoriques. En effet, en partant des équations d'Euler pleinement compressibles dans leur formalisme général en η, les équations pour la coordonnée σ peuvent être obtenues en définissant les fonctions arbitraires $A(\eta)$ et $B(\eta)$ de la manière suivante :

$$\forall \eta \in [0;1], \qquad A(\eta) = 0, \quad B(\eta) = \eta = \sigma \tag{I.7}$$

Dans toutes les discussions à venir, nous pouvons utiliser l'hypothèse de l'isotropie des directions horizontales pour nous permettre de définir ce système linéaire uniquement dans un plan vertical. Ainsi, le système est supposé invariant dans la direction y ($\partial_y \equiv 0$). Les analyses seront donc effectuées dans le plan vertical $(x-\sigma)$ supposé périodique dans la direction x. La procédure permettant la mise en place de la méthode d'analyse précédemment développée pour le système d'Euler défini pour le jeu de variables pronostiques $X = (u, w, T, q, \pi_s)$, est décrite ci-après.

Linéarisation du système borné

La linéarisation s'opère autour de l'état de référence $\overline{X}$ défini comme étant au repos ($\overline{u} = \overline{w} = \overline{\dot{\sigma}} = 0$), isotherme ($\overline{T}$ constant), en équilibre hydrostatique ($\overline{q} = 0$), horizontalement homogène et sans orographie. Après quelques manipulations algébriques, le système d'Euler linéarisé $\mathscr{L}$ apparaît sous la forme :

$$\partial_t u + R\mathcal{G}\nabla T + R\overline{T}\left(\mathcal{I} - \mathcal{G}\right)\nabla q + R\overline{T}\frac{\nabla \pi_s}{\overline{\pi_s}} = 0 \tag{L1}$$

$$\partial_t w - g(\tilde{\partial} + \mathcal{I})q = 0 \tag{L2}$$

$$\partial_t T + \frac{R\overline{T}}{C_v}\left(\nabla u - \frac{1}{\overline{\overline{H}}}\tilde{\partial} w\right) = 0 \tag{L3}$$

$$\partial_t q - \mathcal{S}\nabla u + \frac{C_p}{C_v}\left(\nabla u - \frac{1}{\overline{\overline{H}}}\tilde{\partial} w\right) = 0 \tag{L4}$$

$$\partial_t \pi_s + \overline{\pi_s}\,\mathcal{N}\nabla u = 0 \tag{L5}$$

où $\overline{\pi}_s = p_{00}$, et $\overline{H} = R\overline{T}/g$ correspond à la hauteur d'échelle caractéristique de l'atmosphère de référence, typiquement $\overline{H} \approx 9\,\mathrm{km}$ pour $\overline{T} = 300\,\mathrm{K}$. De plus, $\mathcal{I}$ représente l'opérateur vertical identité tel que $\mathcal{I} \cdot X = X$, et $\tilde{\partial}$, $\mathcal{S}$, $\mathcal{N}$ et $\mathcal{G}$ sont des opérateurs verticaux intervenant de manière usuelle dans la dynamique en coordonnée masse et dont les définitions, dans le cas particulier de

la coordonnée masse σ, sont respectivement données par les relations :

$$\tilde{\partial} X = \sigma \partial_\sigma X \tag{I.8}$$

$$\mathcal{S} X = \frac{1}{\sigma} \int_0^\sigma X \mathrm{d}\sigma' \tag{I.9}$$

$$\mathcal{N} X = \int_0^1 X \mathrm{d}\sigma' \tag{I.10}$$

$$\mathcal{G} X = \int_\sigma^1 \frac{X}{\sigma'} \mathrm{d}\sigma' \tag{I.11}$$

Afin de déterminer les solutions ondulatoires admissibles de ce système, on cherche à obtenir l'équation de structure par élimination successives des variables pour finalement aboutir à une équation pour une variable, en l'occurrence ici w. Ainsi, en dérivant les équations du mouvement (L1) et (L2) par rapport au temps et en procédant par substitution, il vient :

$$\left\{ -\frac{1}{\overline{c}_s^2} \partial_t^4 + \left[\nabla^2 + \frac{1}{\overline{H}^2} \mathcal{L}_v \right] \partial_t^2 + \frac{g^2}{\overline{c}_s^2} \mathcal{C}_2 \nabla^2 \right\} \tilde{\partial} w = g^2 \overline{H} \mathcal{L}_v \left(\frac{C_p}{C_v} \mathcal{I} - \mathcal{S} \right) \nabla^2 \mathcal{C}_1 (\nabla u) \tag{I.12}$$

avec $\mathcal{L}_v = \tilde{\partial}(\tilde{\partial} + \mathcal{I})$, et

$$\mathcal{C}_1 = \mathcal{G}\mathcal{S} - \mathcal{G} - \mathcal{S} + \mathcal{N} \tag{I.13}$$

$$\mathcal{C}_2 = \mathcal{L}_v \left[\mathcal{S}\mathcal{G} - \frac{C_p}{C_v} (\mathcal{G} + \mathcal{S}) \right] \tag{I.14}$$

En utilisant la règle d'intégration par parties, il est aisé de démontrer que $\mathcal{C}_1 \equiv 0$ et $\mathcal{C}_2 \equiv (R/C_v)\mathcal{I}$. Ces deux identités sont également utilisées pour définir les opérateurs discrets. Finalement après quelques manipulations algébriques on obtient l'équation de structure :

$$\left\{ -\frac{1}{\overline{c}_s^2} \partial_t^4 + \left[\nabla^2 + \frac{1}{\overline{H}^2} \mathcal{L}_v \right] \partial_t^2 + \overline{N}^2 \nabla^2 \right\} \tilde{\partial} w = 0 \tag{S}$$

où $\overline{c}_s^2 = R\overline{T}(C_p/C_v)$ est le carré de la vitesse du son, et $\overline{N}^2 = g^2/(C_p\overline{T})$ est le carré de la fréquence de Brunt [4]-Väisälä [5] de l'atmosphère de base, typiquement $\overline{c}_s \approx 350\,\mathrm{m.s^{-1}}$ et $\overline{N} \approx 0{,}018\,\mathrm{s^{-1}}$, pour une température $\overline{T} = 300\,\mathrm{K}$.

L'équation de structure (S) caractérise, comme son nom l'indique, la structure spatio-temporelle des modes propres du système. Les modes propres ondulatoires solutions de l'équation de structure sont recherchés sous la forme $X(x,\sigma,t) = \hat{X}_0\, e^{-\hat{\imath}\omega t} f(x,\sigma)$ avec une fréquence ω et une structure spatiale $f(x,\sigma) = e^{\hat{\imath}kx}\sigma^{\hat{\imath}\hat{\ell}}$, où le nombre d'ondes horizontales k est réel, et le nombre d'ondes verticales $\hat{\ell}$ est supposé complexe. En injectant cette forme dans (S), il apparaît que le caractère oscillatoire ($\omega \in \mathbb{R}$) de ces modes est maintenu si, et seulement si, $\hat{\ell} = \ell + \hat{\imath}/2$, avec ℓ le nombre

4. Sir David Brunt (1886-1965) : météorologue gallois
5. Vilho Väisälä (1889-1969) : météorologue et physicien finlandais

d'ondes verticales réel. Par conséquent, la structure spatiale $f(r)$ des modes propres ondulatoires du système $\mathscr{L}$ est de la forme $f(x,\sigma) = e^{\hat{\imath}kx}\sigma^{\hat{\imath}\ell-1/2}$. Ces modes propres $X = \hat{X}f(x,\sigma)$ satisfont bien $\mathscr{L}X = -\hat{\imath}\omega X$, avec une fréquence réelle ω vérifiant la relation de dispersion :

$$\omega^4 - \overline{c}_s^2\left(k^2 + \frac{\ell^2+1/4}{\overline{H}^2}\right)\omega^2 + \overline{c}_s^2\overline{N}^2k^2 = 0 \tag{I.15}$$

Construction du système non-borné

Pour construire un système linéaire non borné, il faut éliminer π_s, ainsi que les opérateurs $\mathcal{G}$, $\mathcal{S}$, $\mathcal{N}$ qui font intervenir les bords du domaine. Ces opérateurs ne dépendent pas de l'état de référence et surtout satisfont les identités remarquables suivantes :

$$\tilde{\partial}\,\mathcal{G} = -\mathcal{I} \tag{I.16}$$

$$(\tilde{\partial} + \mathcal{I})\mathcal{S} = \mathcal{I} \tag{I.17}$$

$$\tilde{\partial}\,\mathcal{N} = 0 \tag{I.18}$$

Ces identités, qui pour rappel, ne sont valides que pour le système continu dans l'espace, indiquent qu'il convient d'appliquer l'opérateur $\tilde{\partial}$ respectivement aux équations pronostiques de u (L1) et de π_s (L5), et par ailleurs, d'appliquer $(\tilde{\partial}+\mathcal{I})$ à l'équation pronostique en q (L4), pour faire disparaître les termes intégraux du système. Par conséquent, il apparaît de choisir pour opérateur l :

$$l \equiv \begin{pmatrix} \tilde{\partial} \\ \mathcal{I} \\ \mathcal{I} \\ \tilde{\partial} + \mathcal{I} \\ \tilde{\partial} \end{pmatrix}$$

Le système linéarisé non borné $l\mathscr{L}$ s'écrit :

$$\partial_t\tilde{\partial}u - R\nabla T + R\overline{T}\left(\tilde{\partial} + \mathcal{I}\right)\nabla q = 0 \tag{lL1}$$

$$\partial_t w - g(\tilde{\partial} + \mathcal{I})q = 0 \tag{lL2}$$

$$\partial_t T + \frac{R\overline{T}}{C_v}\left(\nabla u - \frac{1}{\overline{H}}\tilde{\partial}w\right) = 0 \tag{lL3}$$

$$\partial_t\left(\tilde{\partial} + \mathcal{I}\right)q - \nabla u + \frac{C_p}{C_v}\left(\tilde{\partial} + \mathcal{I}\right)\left(\nabla u - \frac{1}{\overline{H}}\tilde{\partial}w\right) = 0 \tag{lL4}$$

Comme la condition [C1] est bien réalisée, ce système ne fait donc aucune allusion aux conditions aux bords. En procédant également par élimination comme dans le cas borné, on peut facilement vérifier que l'on obtient exactement la même équation de structure (S). Par conséquent, les systèmes borné et non-borné ont les mêmes modes normaux $\hat{X}(t)f(r)$, de structure spatiale identique de

la forme $f(r) = e^{\hat{\imath}kx}\sigma^{\hat{\imath}\ell-1/2}$ et de fréquence égale ω vérifiant la relation de dispersion (I.15). La condition [C2] est donc vérifiée.
Par ailleurs, pour ces modes, on a $l_i f(r) = \xi_i f(r)$ avec :

$$\begin{aligned}
\xi_1 &= \left(\hat{\imath}\ell - \frac{1}{2}\right) \\
\xi_2 &= 1 \\
\xi_3 &= 1 \\
\xi_4 &= \left(\hat{\imath}\ell + \frac{1}{2}\right)
\end{aligned}$$

avec $\xi_5 = \xi_1$. Ainsi, dans l'espace des modes propres, le système non-borné s'écrit :

$$\xi_1 \partial_t \hat{u} = \mu_{13}\,\hat{T} + \mu_{14}\,\hat{q} \tag{I.19}$$

$$\xi_2 \partial_t \hat{w} = \mu_{24}\,\hat{q} \tag{I.20}$$

$$\xi_3 \partial_t \hat{T} = \mu_{31}\,\hat{u} + \mu_{32}\,\hat{w} \tag{I.21}$$

$$\xi_4 \partial_t \hat{q} = \mu_{41}\,\hat{u} + \mu_{42}\,\hat{w} \tag{I.22}$$

avec :

$$\begin{aligned}
\mu_{13} &= R\hat{\imath}k \\
\mu_{14} &= -R\overline{T}\hat{\imath}k\left(\hat{\imath}\ell + \frac{1}{2}\right) \\
\mu_{24} &= g\left(\hat{\imath}\ell + \frac{1}{2}\right) \\
\mu_{31} &= -\frac{R\overline{T}}{C_v}\hat{\imath}k \\
\mu_{32} &= \frac{g}{C_v}\left(\hat{\imath}\ell - \frac{1}{2}\right) \\
\mu_{41} &= \hat{\imath}k - \frac{C_p}{C_v}\hat{\imath}k\left(\hat{\imath}\ell + \frac{1}{2}\right) \\
\mu_{42} &= \frac{C_p}{C_v}\frac{1}{\overline{H}}\left(\hat{\imath}\ell + \frac{1}{2}\right)\left(\hat{\imath}\ell - \frac{1}{2}\right)
\end{aligned}$$

Les conditions [C3]-[C4] sont par conséquent réalisées.

Finalement, la stabilité du système non-borné discrétisé en temps peut donc être analysée selon le principe de Von Neumann, en utilisant la forme scalaire du système non-borné défini par les équations (I.19)-(I.22) ci-dessus.

3 Paramètres de l'analyse pour un schéma HEVI

Revenons sur les solutions ondulatoires du système borné (qui sont aussi solutions du système non-borné). Pour un nombre d'ondes horizontales k, et un nombre d'ondes verticales ℓ donnés, la relation de dispersion (I.15) admet quatre racines en ω. Chaque paire de racines a des signes opposés tels que $\omega \in \{\pm\omega_a; \pm\omega_g\}$. Bien que les ondes rapides supportées par ce système soient de nature mixtes (*ie :* à la fois acoustique et de gravité), nous désignerons par ω_a la fréquence des modes acoustiques qui, par définition, est plus élevée que ω_g, celle des modes de gravité. Chacune de ces fréquences est donnée par :

$$\omega_a^{\pm} = \pm\overline{c}_s \left(k^2 + \frac{\ell^2 + 1/4}{\overline{H}^2} \right)^{1/2} \left\{ \frac{1}{2} + \left[\frac{1}{4} - \frac{\overline{N}^2 k^2}{\overline{c}_s^2 \left(k^2 + \frac{\ell^2 + 1/4}{\overline{H}^2} \right)^2} \right]^{1/2} \right\}^{1/2} \tag{I.23}$$

$$\omega_g^{\pm} = \pm\overline{c}_s \left(k^2 + \frac{\ell^2 + 1/4}{\overline{H}^2} \right)^{1/2} \left\{ \frac{1}{2} - \left[\frac{1}{4} - \frac{\overline{N}^2 k^2}{\overline{c}_s^2 \left(k^2 + \frac{\ell^2 + 1/4}{\overline{H}^2} \right)^2} \right]^{1/2} \right\}^{1/2} \tag{I.24}$$

Les relations ci-dessus permettent de vérifier que les modes acoustiques se propagent effectivement à des fréquences bien plus rapides que celles des modes de gravité ($\omega_a > \omega_g$). En conséquence, pour un pas de temps Δt, le nombre de Courant ondulatoire (par opposition au nombre de Courant advectif imposé par le transport par le vent) le plus contraignant est celui imposé par les modes acoustiques. Il est noté $C_a = \omega_a \Delta t$. Selon l'approche HEVI, on distingue la contribution des termes d'ajustement verticaux associés au nombre de Courant vertical $C_{a,z}$, lesquels subissent un traitement implicite, et la contribution des termes d'ajustement horizontaux associés au nombre de courant horizontal $C_{a,x}$ qui, eux, seront traités explicitement. Dans la mesure où le dernier terme entre crochet de l'équation (I.23) est au plus égal à 1, les nombres de Courant $C_{a,z}$ et $C_{a,x}$ sont respectivement définis comme

$$C_{a,z} = \frac{\overline{c}_s}{\overline{H}} \Delta t \sqrt{\ell^2 + 1/4}, \qquad \text{et} \qquad C_{a,x} = \overline{c}_s k \Delta t \tag{I.25}$$

Sur une grille horizontale régulière, la valeur maximale de $C_{a,x}$ est atteinte pour la plus petite longueur d'onde pouvant être résolue par le modèle c'est-à-dire $2\Delta x$, correspondant à un nombre d'onde maximum $k_{\max} = \pi/\Delta x$. Sachant que les petites longueurs d'onde du modèle sont généralement celles qui engendrent le plus de problèmes numériques, l'analyse se focalisera sur le nombre de Courant horizontal maximum défini comme :

$$C_* = \overline{c}_s k_{\max} \Delta t = \overline{c}_s \frac{\pi}{\Delta x} \Delta t \tag{I.26}$$

À partir de la connaissance du nombre d'onde maximum $k_{\max}$ résolu par le modèle, ainsi que de la valeur limite du nombre de Courant horizontal $C_{*\text{limite}}$ autorisée par le schéma HEVI, nous

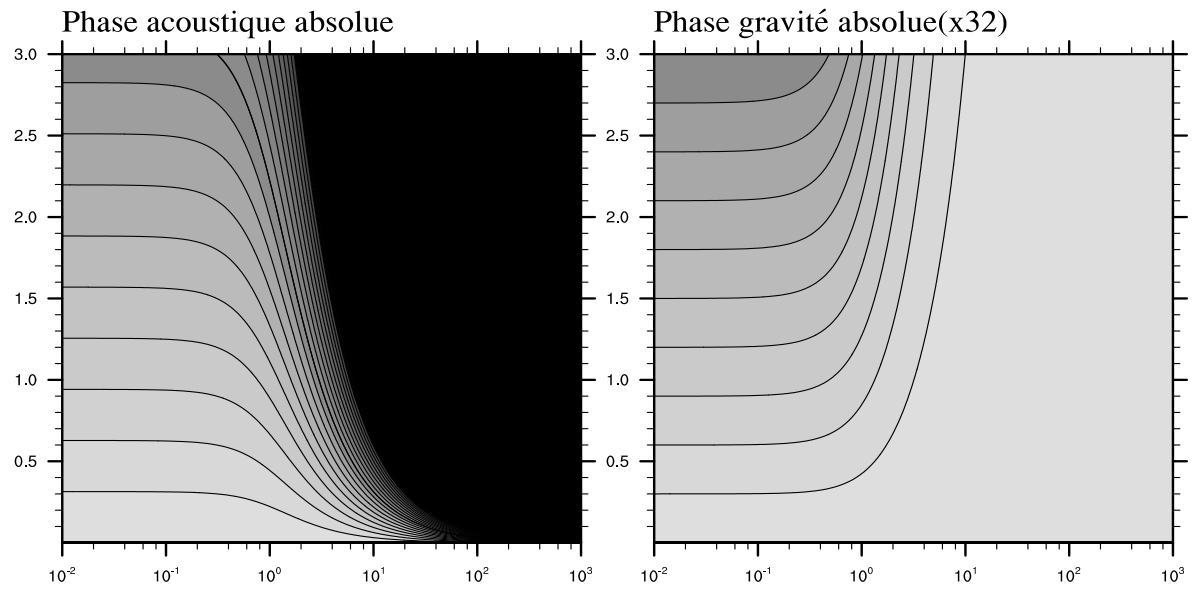


FIGURE I.1 – *Amplitude des phases exactes absolues des ondes rapides du système en fonction de C_* (ordonnée) et de r (abscisse) pour un état de référence repos. Le panneau de gauche représente la phase absolue des ondes acoustiques alors que le panneau de droite représente la phase absolue (multiplié par 32) des ondes de gravité.*

pouvons facilement déduire un pas de temps critique $\Delta t_{\text{limite}} = (C_{*\text{limite}}/\overline{c}_s k_{max})$. En fait, bien que le nombre de Courant limite ne dépende que du choix du schéma de discrétisation temporelle, le pas de temps critique est, quant à lui, également dépendant du choix de la discrétisation spatiale.

En effet, la définition du nombre de Courant ondulatoire horizontal C_* ci-dessus, suppose implicitement que la discrétisation horizontale du modèle est spectrale ; discrétisation pour laquelle la fonction de réponse spectrale de l'opérateur Laplacien [6] horizontal est $\widehat{\partial_x \circ \partial_x}(k) \to -k^2$, pour $k \in [0; k_{\max}]$. Pour un modèle employant des méthodes de discrétisation horizontale point-de-grille (*ie :* différences, volumes ou éléments finis) sur des grilles régulières (ou subissant une transformation conforme), les fonctions de réponse sont différentes de celles de la méthode spectrale. Dans ce contexte, en notant $\mathrm{d}(k) = \sqrt{\widehat{\partial_x \circ \partial_x}(k)}$ la racine carrée de la fonction de réponse spectrale de l'opérateur Laplacien horizontal discret $\partial_x \circ \partial_x$, résultant de la composition des opérateurs discrets de divergence horizontale (*ie :* $\partial_x u$) et de gradient horizontal de pression (*ie :* $\partial_x q$), le nombre de Courant point-de-grille, noté $C_{\mathrm{GP}*}$, par opposition au nombre de Courant spectral C_* défini plus haut, doit en toute rigueur être défini comme $C_{\mathrm{GP}*} = \overline{c}_s \mathrm{d}_{\max} \Delta t$, avec $\mathrm{d}_{\max} = \mathrm{d}(k_{\max}) \leq k_{\max}$. Cette dernière inégalité traduit le simple fait que la discrétisation spectrale fournit la meilleure évaluation possible des dérivées horizontales sur un domaine périodique et pour des champs supposés relativement réguliers. Finalement, pour un nombre de Courant limite $C_{*\text{limite}}$ le pas de temps critique est en toute rigueur donné par $\Delta t_{\text{limite}} = (C_{*\text{limite}}/\overline{c}_s \mathrm{d}(k_{\max}))$, où $\mathrm{d}(k_{\max})$ est à déterminer en fonction de la géométrie et de la discrétisation horizontale du modèle.

6. Pierre-Simon de Laplace (1749-1827) : mathématicien, astronome, physicien et homme politique français

Sur la verticale, le nombre de Courant vertical $C_{a,z}$ est lié au nombre de Courant horizontal C_* via le paramètre r défini comme le rapport :

$$r = \frac{C_{a,z}}{C_*} = \frac{1}{\overline{H}} \left[\frac{\ell^2}{k_{\max}^2} + \frac{1}{4k_{\max}^2} \right]^{1/2} \approx \frac{|\ell| \Delta x}{\pi \overline{H}} \tag{I.27}$$

L'approximation ci-dessus est vérifiée pour des nombres d'ondes verticales ℓ non-identiquement nuls, ainsi que pour un toit du modèle placé à une altitude H_T bien en deçà de $4\pi\overline{H} \approx 115\,\mathrm{km}$ (ce qui est généralement le cas). Le paramètre r représente un rapport d'aspect dont les valeurs minimale et maximale dépendent des caractéristiques dimensionnelles du modèle. La plus petite longueur d'onde résolue sur la verticale est telle que $\ell_{\max}/\overline{H} = \pi/\Delta z$, avec des valeurs de Δz pouvant variées entre 10 m (à la surface du modèle afin de mieux capturer les processus de couches limites), à 1 km au sommet du modèle situé à environ 30 km. Sur l'horizontale, les résolutions cibles actuelles en PNT sont typiquement de l'ordre du kilomètre $\Delta x = 1\,\mathrm{km}$, et peuvent atteindre 10 km pour les applications climatiques. Par conséquent, le rapport d'aspect r varie approximativement dans l'intervalle $[10^{-2}, 10^3]$. Quant au nombre de Courant ondulatoire horizontal C_*, il est admis que dans la mesure où les schémas HEVI étudiés ici traitent l'horizontale de façon explicite, la valeur limite autorisée pour C_* n'excède pas $2\sqrt{2}$ correspondant au nombre de Courant limite d'un schéma Runge-Kutta d'ordre 4 purement explicite. D'où finalement, en arrondissant à la valeur entière supérieure nous prenons C_* dans l'intervalle $[0, 3]$.

Les phases exactes au bout d'un intervalle Δt des modes acoustiques et gravité de longueur d'onde horizontal $2\Delta x$ peuvent s'écrire en fonction des paramètres C_*, r et $k_{\max}$ de la manière suivante :

$$\theta_a = \omega_a^{\pm} \Delta t = \pm C_* \left(1 + r^2\right)^{1/2} \left\{ \frac{1}{2} + \left[\frac{1}{4} - \frac{\overline{N}^2/\overline{c}_s^2}{k_{\max}^2 \left(1 + r^2\right)^2} \right]^{1/2} \right\}^{1/2} \tag{I.28}$$

$$\theta_g = \omega_g^{\pm} \Delta t = \pm C_* \left(1 + r^2\right)^{1/2} \left\{ \frac{1}{2} - \left[\frac{1}{4} - \frac{\overline{N}^2/\overline{c}_s^2}{k_{\max}^2 \left(1 + r^2\right)^2} \right]^{1/2} \right\}^{1/2} \tag{I.29}$$

La figure (I.1) représente les amplitudes des phases θ_a (acoustiques) et θ_g (gravité) en absence d'écoulement de l'état de référence, en fonction du nombre de Courant horizontal C_* (en ordonnée) et du rapport d'aspect r (en abscisse), pour $\Delta x = 1\,\mathrm{km}$, soit $k_{\max} \approx 0{,}00314\,\mathrm{m}^{-1}$. Les zones noires résultent de l'accumulation d'un nombre très important de contours. Les phases des modes de gravité ont été multipliées par 32 afin d'assurer une meilleure comparaison visuelle. Les phases numériques calculées dans ce travail devront être comparées à ces phases exactes afin de déterminer, l'erreur de phase commise du fait de la discrétisation temporelle.

Il est à noter qu'en présence d'un écoulement de base horizontal $\overline{U}$, nous définirons le nombre de Mach $M_U = |\overline{U}|/\overline{c}_s$ pour quantifier les effets de l'advection sur le schéma. Dans la littérature, il est d'usage commun d'utiliser le nombre de Courant advectif à cet effet, lequel serait défini dans ce contexte HEVI comme $C_U^* = |\overline{U}| k_{\max} \Delta t$. Cependant, il existe un lien direct entre ce nombre de Courant advectif horizontal, le nombre de Mach et le nombre de Courant ondulatoire horizontal, tel que $C_U^* = M_U C_*$. Dans la mesure où certains jets d'altitude peuvent atteindre des vitesses

allant jusqu'à $900\,\mathrm{km.h^{-1}}$ en particulier dans la haute stratosphère $H_T \gtrsim 30\,\mathrm{km}$, les valeurs de M_U sont susceptibles de varier dans l'intervalle $[0, 1]$.

Finalement, les paramètres pertinents choisis pour l'analyse numérique des schémas temporels étudiés dans ce travail sont : le nombre de Courant ondulatoire horizontal C_*, le rapport d'aspect r, et le nombre de Mach horizontal M_U.

4 Pertinence de l'analyse linéaire continue en espace

La méthode d'analyse numérique continue en espace proposée ici repose sur l'hypothèse d'une évolution linéaire des variables perturbées autour d'un état de référence très simple. À l'utilisation d'une telle approche, on pourrait raisonnablement objecter que l'atmosphère réelle est bien plus complexe, notamment le fait qu'elle est assez loin d'être isotherme. Et que seule la validation expérimentale à l'aide d'un modèle non-linéaire pour des conditions atmosphériques réalistes permet d'établir les véritables propriétés de stabilité des schémas.

Toutefois, il est un fait que l'approche linéaire nous enseigne : si le schéma de discrétisation temporelle s'avère instable dans ce contexte linéaire simplifié, alors les chances que ce schéma puisse être utilisable dans le cadre d'un modèle non-linéaire opérationnel sont extrêmement minces. En d'autres termes, la méthode d'analyse que nous proposons doit être appréhendée comme un premier outil d'investigation permettant :

- de discriminer les méthodes de discrétisation temporelle entre elles, en mettant en évidence le bénéfice potentiel d'un choix de schéma particulier par rapport à un autre.
- écarter les schémas ne présentant pas les propriétés de stabilité et de précision requises pour des applications PNT.
- d'élaborer des schémas temporels plus stables.

Une autre faiblesse inhérente à l'analyse continue en espace est qu'elle ne tient pas compte des effets des bords du domaine. En fait, la présence des bords réduit le nombre de degrés de liberté sur la verticale, n'autorisant que les solutions qui satisfont les conditions aux bords. Dans ces circonstances, il est fort probable que les effets des bords soient plus de nature stabilisatrice que déstabilisatrice, et donc que l'analyse continue en espace puisse, de ce fait, surestimer les taux de croissance (en présence d'instabilités numériques). Afin d'y remédier une analyse linéaire discrète sur la verticale a également été mise en place dans ce travail. Elle consiste à représenter les variables atmosphériques (de l'état référence et de l'état perturbé) comme des vecteurs colonnes dont les dimensions sont égales au nombre de niveaux verticaux du modèle. Cette analyse discrète permet de confirmer ou d'infirmer les résultats de l'analyse continue dans un cadre un peu plus proche de celui du vrai modèle numérique.

Chapitre II

Méthode du pas de temps fractionné sous la contrainte HEVI

Les modèles décrivant la dynamique du temps supportent plusieurs phénomènes ayant chacun sa vitesse de propagation propre. Dans le cas des équations dites « *raides* », ces vitesses sont, par définition, très différentes selon que l'on considère les processus les plus lents et les plus rapides. C'est la raison pour laquelle des méthodes implicites sont, pour ces cas, largement utilisées pour s'affranchir de la plus forte contrainte CFL. Dans ce travail, le système que nous tentons de résoudre est celui des équations d'Euler non-hydrostatiques pleinement compressibles qui peut s'écrire formellement :

$$\partial_t X = \mathscr{M}(X) = \mathscr{S}(X) + \mathscr{F}(X) \tag{II.1}$$

où $\mathscr{S}$ représente l'opérateur non-linéaire décrivant les processus désignés comme relativement lents et $\mathscr{F}$ est l'opérateur non-linéaire associé aux processus les plus rapides.

Dans un contexte opérationnel, où l'économie des calculs est une contrainte continuelle, il est parfois envisageable d'utiliser des schémas spécifiques pour des processus ayant des vitesses de propagation très différentes. Si des méthodes explicites sont utilisées, alors, à cause de la contrainte de stabilité CFL, le schéma réalisant l'intégration des processus les plus rapides conditionne un pas de temps relativement plus petit que le schéma intégrant les processus plus lents. Ainsi, par essence, les processus relativement plus lents n'évoluent que faiblement pendant la durée décrite par le petit pas de temps. Par conséquent, il est possible de considérer ces phénomènes comme stationnaires durant une période couvrant plusieurs de ces pas de temps. Numériquement, cela permet de maintenir constants les termes traitant la partie lente durant plusieurs itérations sur les petits pas de temps, et ainsi d'économiser les calculs de l'évaluation des termes lents.

Afin d'être capable d'appliquer cette méthode pour le système d'Euler, il faut, en amont, être capable de répondre à plusieurs questions. Quels sont les phénomènes ayant la plus grande vitesse de propagation, et quels sont les termes qui les modélisent ? Afin de respecter la contrainte HEVI, sans affecter la structure des solutions du système, quelles propriétés doivent respecter le schéma intégrant dans le temps ces termes rapides ? Enfin, comment traiter les autres termes, et avec quel schéma ?

Pour répondre à ces questions, ce chapitre va présenter une nouvelle approche pour adapter ces méthodes aux équations d'Euler pleinement compressibles en coordonnée masse. De plus, un nouveau schéma d'intégration de la partie lente, qui augmente considérablement la stabilité de la méthode sans augmenter de manière trop importante le nombre d'itérations, sera proposé. Pour étayer cette réflexion, la démarche consistera à présenter de manière très générale ces méthodes. Les différentes contraintes liées à la PNT guideront les choix qui imposent de sélectionner uniquement certains schémas. Ce contexte bien établi, nous pourrons exposer des idées d'amélioration, et confirmer leur efficacité par des études de stabilité dans un cadre simplifié.

1 Présentation de la méthode

Les méthodes HEVI, faisant l'hypothèse de séparabilité des processus, se divisent en deux familles. La première est le *additive-splitting* proposée par Marchuk (1974) [45] et analysée par LeVeque & Oliger (1983) [40]. Elles visent à intégrer séparément d'une part les processus rapides, et d'autre part, les processus lents. Soit Δt le pas de temps réalisant l'intégration des termes responsables des processus lents, et soient $\Phi_f(\Delta t)$ et $\Phi_s(\Delta t)$ les schémas intégrant respectivement les processus rapides et lents. Alors que le système s'écrit comme la somme de ces deux parties, les méthodes additive-splitting calculent un premier état intermédiaire issu uniquement de l'intégration des termes rapides par $\Phi_f(\Delta t)$. Comme ce schéma possède une plus forte contrainte sur la stabilité, le pas de temps est divisé en M sous-pas de temps, de sorte que le schéma réalisant l'intégration de ces processus s'écrit $\Phi_f^M(\Delta t/M)$. Puis, l'évolution des termes lents est évaluée à partir de cet état intermédiaire par $\Phi_s(\Delta t)$. Au final, les schémas utilisant la méthode additive-splitting s'écrivent $\Phi_s(\Delta t) \circ \Phi_f^M(\Delta t/M)$. Cette méthode souffre de plusieurs problèmes. Le premier porte sur la précision, car malgré l'ordre des schémas $\Phi_f(\Delta t)$ et $\Phi_s(\Delta t)$, la méthode décrite ci-avant reste toujours d'ordre 1 en temps. Ce problème a été partiellement résolu par Stang (1968) [65] qui propose une seconde étape avec le schéma $\Phi_f^{M/2}(\Delta t/2M) \circ \Phi_s(\Delta t) \circ \Phi_f^{M/2}(\Delta t/2M)$. Néanmoins Purser & Leslie (1991) [53] montrent que les erreurs restent toujours importantes. Le second problème de ces méthodes porte sur la stabilité. En effet, LeVeque & Oliger (1983) [40] démontrent que, si les deux schémas sont stables pour le même pas de temps, alors la méthode globale est stable si, et seulement si, les opérateurs modélisant les deux parties commutent. Cette contrainte explique pourquoi ces méthodes n'ont jamais été utilisées pour résoudre le système d'Euler non-hydrostatique, bien qu'elles aient été testées avec succès pour le modèle des équations hydrostatiques. La difficulté de l'utilisation de ces méthodes oriente donc vers une intégration simultanée des deux parties. Ce sont alors les schémas *au pas de temps fractionné*.

L'idée des méthodes fractionnant le pas de temps est de traiter de manière différenciée la partie du modèle traitant les processus rapides et le reste du modèle. Les méthodes SI peuvent donc être décrites par ce formalisme, mais la condition HEVI impose, par la contrainte CFL, un pas de temps critique pour le schéma des processus rapides $\Delta\tau$ et un pas de temps relativement plus grand $\Delta t = M\Delta\tau$ pour le schéma portant sur la partie lente. L'idée est donc d'intégrer l'ensemble des équations sur des petits pas de temps ; les termes responsables des processus rapides sont calculés à chaque itération, alors que les termes responsables de l'évolution des processus lents seront constants et traités comme des termes sources durant un certain nombre d'itérations. La

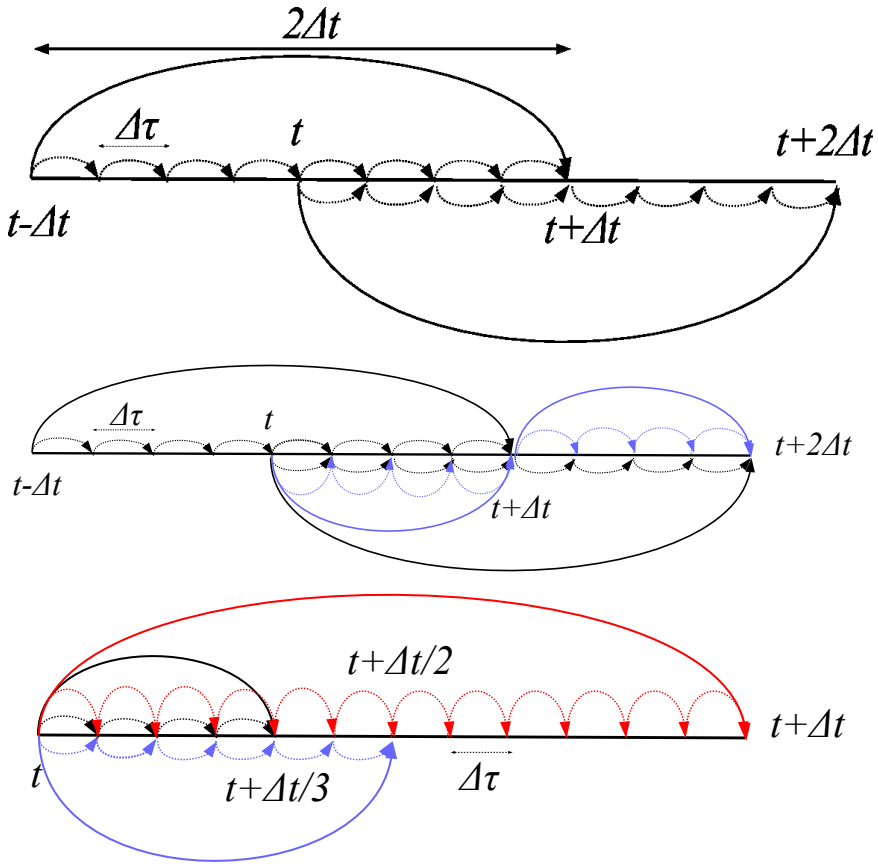


FIGURE II.1 – *Principe du temps fractionné avec un schéma saute-mouton (en haut), Kurihara (au milieu) et Runge-Kutta-3 (en bas).*

durée pendant laquelle les termes sources sont constants ainsi que l'initialisation de l'algorithme sur les sous-pas de temps, doivent respecter la consistance du schéma d'intégration des processus lents. La FigII.1 illustre ce principe de sous-pas de temps et de mise à jour des termes sources afin de construire un schéma d'intégration temporelle pour les termes advectifs. Introduites par Klemp & Wilhelmson (1978) [35] pour simuler les tempêtes, et après l'ajout d'études sur leurs stabilités et leurs précisions pour le système non-hydrostatique (Skamarock & Klemp (1992, 1994) [60, 61]) ces méthodes se sont propagées à de nombreux modèles de prévision notamment grâce à leur simplicité d'application.

Comme le pas de temps du modèle se définit par rapport à la partie du système modélisant la propagation des phénomènes les plus rapides, il est important de commencer par définir cette partie, et le schéma qui l'intègre.

2 Construction de l'opérateur associé aux processus rapides

Commençons par étudier l'opérateur qui définit le pas de temps nominal pour l'intégration numérique du système. Dans un premier temps, étudions les propriétés que doit respecter cet opérateur pour pouvoir le définir. Ensuite, les contraintes opérationnelles guideront les propriétés que doit posséder le schéma intégrant ces termes. Ceci impose une première condition sur la stabilité globale de la méthode.

Les phénomènes ayant la plus grande vitesse de propagation sont les ondes élastiques. Or, chaque onde est le produit d'une perturbation autour d'un équilibre stable. C'est pourquoi, les processus rapides sont représentés par un système linéaire. D'un point de vue numérique, cette propriété permet d'envisager un traitement implicite car le problème inverse, qui est généré par ce traitement, est largement plus aisé si l'équation est linéaire. Formellement, ce système apparaît par le développement limité du modèle complet $\mathscr{M}$ autour d'un état de référence X^r considéré au repos :

$$\mathscr{M}(X) \approx \mathscr{M}(X^r) + \mathscr{L}^r \cdot (X - X^r)$$

où $\mathscr{L}^r$ est l'opérateur $\mathscr{M}$ linéarisé autour de X^r. D'après les théories d'analyses mathématiques, nous savons que cette approximation reste viable tant que les variables X et X^r sont suffisamment proches (au sens des distributions). De là, l'identification avec l'équation (II.1) définit les opérateurs $\mathscr{S}$ et $\mathscr{F}$ suivants :

$$\mathscr{S}(X) = (\mathscr{M} - \mathscr{L}^r)(X) \tag{II.2}$$
$$\mathscr{F}(X) = \mathscr{L}^r \cdot X \tag{II.3}$$

L'efficacité de la méthode repose donc sur le fait que l'opérateur $\mathscr{S}$ traite, le plus possible, uniquement les processus lents. Ceci est donc équivalent, par le développement limité, à ce que l'état pris comme la référence soit le plus proche possible de l'état courant. Dans le cas où ces deux états sont trop éloignés, il risque donc de subsiter des termes régissant l'évolution d'ondes rapides dans la partie traitée avec un grand pas de temps. Dans un tel cas, ces instabilités non-linéaires peuvent rendre ces méthodes inutilisables. Afin d'assurer que de telles instabilités n'apparaissent, les méthodes initialement mises en œuvre par Klemp & Wilhelmson (1978) [35] considéraient que la partie lente était principalement composée des processus advectifs, sans la soustraction de la partie linéaire, et le système linéaire supposait que l'instant courant t était en équilibre hydrostatique et uniforme. Ces hypothèses semblent donc ne pas suivre le formalisme proposé ci-avant, car si la différence entre le modèle complet $\mathscr{M}$ et le système linéaire $\mathscr{L}^r$ était réduite à l'advection seule, alors le système linéaire ne pourrait pas respecter toutes les hypothèses qui le définissent. Il semble donc qu'un nombre important de termes ne soit pas traité par ces méthodes. Par souci de complémentarité, il semble donc préférable d'accepter la présence de résidus non-linéaires dans la partie lente, mais de les contrôler en définissant l'état de référence comme étant le plus proche possible de l'état à l'instant courant.

$$|| \underbrace{\mathscr{M}(X) - \mathscr{L}^t \cdot X}_{\mathscr{S}(X)} || \ll ||\mathscr{L}^t \cdot X||$$

Le système linéaire à utiliser doit cependant respecter certaines contraintes. Comme il est censé modéliser la propagation des ondes rapides, la linéarisation des termes s'effectue uniquement sur les termes soumis à un opérateur de dérivation, et les coefficients de ces opérateurs peuvent être gardés à l'instant t, sans avoir besoin de les linéariser. Par ailleurs, les coefficients calculés par les variables pronostiques (tels que le géopotentiel), ne peuvent donc pas être utilisés de manière implicite. C'est pourquoi, le système linéaire $\mathscr{L}^t$ est de la forme :

$$\partial_t \mathbf{V} + RT^t \left(\frac{\nabla \pi_s}{\pi_s^t} + \nabla q \right) + R\mathcal{G}^t(\mathrm{e}^{-q^t} \nabla T) - R\mathcal{G}^t(T^t \nabla q) +$$
$$R\mathcal{G}^t(T^t \mathrm{e}^{-q^t} \nabla \pi_s) + \nabla \phi^t \cdot \left(\tilde{\partial} + \mathrm{I} \right)^t (\mathrm{e}^{q^t} q) = 0 \tag{II.4}$$
$$\partial_t w - g \left(\tilde{\partial} + \mathrm{I} \right)^t (\mathrm{e}^{q^t} q) = 0 \tag{II.5}$$
$$\partial_t T + \frac{RT^t}{C_v} \left(\nabla \cdot \mathbf{V} - \frac{1}{H^t} \tilde{\partial}^t w + \frac{\nabla \phi^t}{RT^t} \cdot \tilde{\partial}^t \mathbf{V} \right) = 0 \tag{II.6}$$
$$\partial_t q - \mathcal{S}^t(\nabla \cdot \mathbf{V}) + \frac{C_p}{C_v} \left(\nabla \cdot \mathbf{V} - \frac{1}{H^t} \tilde{\partial}^t w + \frac{\nabla \phi^t}{RT^t} \cdot \tilde{\partial}^t \mathbf{V} \right) = 0 \tag{II.7}$$
$$\partial_t \pi_s + \mathcal{N}^t(\nabla \cdot \mathbf{V}) = 0 \tag{II.8}$$

avec $H^t = RT^t/g$, une hauteur caractéristique.

Pour vérifier que ce système modélise bien la propagation des ondes rapides, mais aussi pour savoir comment calculer chaque terme, il faut étudier les caractéristiques que doit respecter le schéma intégrant ce système.

Choix du traitement temporel des processus rapides

Reste à savoir comment évaluer le système linéaire durant les sous-pas de temps. Dans un souci d'efficacité, les seuls algorithmes utilisés pour cette résolution sont des schémas d'Euler explicites sur l'horizontale, ainsi que des méthodes implicites afin de s'affranchir de la condition de stabilité CFL portant sur la propagation verticale des ondes rapides. Ces seules conditions suffisent à lever beaucoup de degrés de liberté. Dans ce qui suit, nous allons omettre la partie lente et commencer l'étude de stabilité du système linéaire $\mathscr{L}$. En notant λ le nombre complexe tel que chaque variable à l'instant $\tau + \Delta\tau$ (notée X^+) s'écrive sous la forme d'une suite géométrique $X^+ = \lambda X^0$ (avec X^0 la même variable à l'instant τ), nous cherchons les conditions que doivent vérifier les différents traitements de sorte que le schéma respecte deux propriétés. La première ; que les solutions numériques aient la même structure spatio-temporelle que les solutions continues. Pour cela, il suffit de reconstituer l'équation de structure discrétisée en temps et de faire émerger une analogie avec l'équation de structure continue (S). La seconde propriété, est que le schéma

global doit être conditionnellement stable. Comme le système est conservatif, cette condition est équivalente à ce que la norme de ce coefficient complexe λ soit inférieure ou égale à 1 ($\Gamma = |\lambda| \leq 1$). Suivant ces définitions, nous introduisons de nouvelles notations pour alléger la présentation :

$$\Lambda_t = (\lambda - 1)/\Delta\tau \tag{II.9}$$

$$\Lambda_{\epsilon_{\varphi_1\varphi_2}} = \epsilon\lambda + (1 - \epsilon) \tag{II.10}$$

Le coefficient Λ_t représente la dérivée temporelle discrète. Un traitement explicite de la variable φ_2 dans l'équation de φ_1 est modélisé avec $\epsilon = 0$, alors qu'un traitement implicite (respectivement trapézoïdal) est obtenu avec $\epsilon = 1$ (respectivement $\epsilon = 1/2$). La discrétisation temporelle du système linéaire $\mathscr{L}$ est donc réduite à :

$$\Lambda_t u + \Lambda_{\epsilon_{vt}}\mathcal{G}\nabla T + \Lambda_{\epsilon_{vq}}R\overline{T}(\mathcal{I} - \mathcal{G})\nabla q + \Lambda_{\epsilon_{vp}}\frac{R\overline{T}}{\overline{\pi_s}}\nabla\pi_s = 0 \tag{II.11}$$

$$\Lambda_t w - \Lambda_{\epsilon_{wq}} g(\tilde{\partial} + \mathcal{I})q = 0 \tag{II.12}$$

$$\Lambda_t T + \frac{R\overline{T}}{C_v}\left(\Lambda_{\epsilon_{tv}}\nabla u - \Lambda_{\epsilon_{tw}}\frac{1}{\overline{\overline{H}}}\tilde{\partial}w\right) = 0 \tag{II.13}$$

$$\Lambda_t q - \Lambda_{\epsilon_{qv}}\overline{S}\nabla u + \frac{C_p}{C_v}\left(\Lambda_{\epsilon_{qv}}\nabla u - \Lambda_{\epsilon_{qw}}\frac{1}{\overline{\overline{H}}}\tilde{\partial}w\right) = 0 \tag{II.14}$$

$$\Lambda_t\pi_s + \Lambda_{\epsilon_{pv}}\overline{\pi_s}\mathcal{N}\nabla u = 0 \tag{II.15}$$

Suivant le même procédé que pour l'établissement de l'équation de structure continue (S), nous allons établir l'équation de structure temporellement discrétisée, et voir les différentes hypothèses sur les traitements des variables nécessaires pour retrouver une structure similaire pour les solutions numériques. Tout d'abord, dérivons une seconde fois les équations du mouvement et substituons les variables de la température et de la pression dans les équations du mouvement (II.11) et (II.12) :

$$\begin{aligned}\Big\{\Lambda_t^2 + R\overline{T}\Big[\left(\Lambda_{\epsilon_{vq}}\Lambda_{\epsilon_{qv}} - \Lambda_{\epsilon_{vt}}\Lambda_{\epsilon_{tv}}\right)\frac{R}{C_v}\mathcal{G} + \left(\Lambda_{\epsilon_{vq}}\Lambda_{\epsilon_{qv}} - \Lambda_{\epsilon_{vp}}\Lambda_{\epsilon_{pv}}\right)\overline{\mathcal{N}} \\ -\Lambda_{\epsilon_{vq}}\Lambda_{\epsilon_{qv}}\frac{C_p}{C_v}\Big]\nabla^2\Big\}u + g\left\{\left[\Lambda_{\epsilon_{vt}}\Lambda_{\epsilon_{tw}}\frac{R}{C_v} - \Lambda_{\epsilon_{vq}}\Lambda_{\epsilon_{qw}}\frac{C_p}{C_v}\right]\mathcal{G} + \Lambda_{\epsilon_{vq}}\Lambda_{\epsilon_{qw}}\frac{C_p}{C_v}\right\}\nabla\tilde{\partial}w = 0\end{aligned} \tag{II.16}$$

$$\left\{\Lambda_t^2 - \Lambda_{\epsilon_{wq}}\Lambda_{\epsilon_{qw}}\frac{\overline{c}_s^2}{\overline{\overline{H}}^2}\mathcal{L}_v\right\}\tilde{\partial}w + g\Lambda_{\epsilon_{wq}}\Lambda_{\epsilon_{qv}}\mathcal{L}_v\left\{-\mathcal{S} + \frac{C_p}{C_v}\right\}\nabla\tilde{\partial}u = 0 \tag{II.17}$$

Pour établir cette équation, il est nécessaire de réécrire la condition portant sur l'opérateur $\mathcal{C}_1$ (*ie :* $\mathcal{C}_1 \equiv 0$), qui définit des relations entre les opérateurs d'intégration verticale :

$$(\mathcal{I} - \mathcal{G}) \cdot \left(\mathcal{S} - \frac{C_p}{C_v}\right) \equiv \mathcal{N} - \frac{R}{C_p}\mathcal{G} - \frac{C_p}{C_v}$$

Cette équation est fondamentale pour l'écriture de l'équation de structure discrète. Elle ne peut être obtenue qu'en imposant un unique traitement de u dans l'équation (II.14). Cette condition impose, pour le modèle non-linéaire, que la divergence horizontale et le terme $\dot{\pi}/\pi$ soient évalués de la même façon. Cette remarque sera importante quand nous parlerons de la mise en œuvre des méthodes HEVI pour le système complet.
Afin de retrouver la forme discrétisée de l'équation de structure, il est donc nécessaire d'avoir :

$$\Lambda_{\epsilon_{vq}}\Lambda_{\epsilon_{qv}} = \Lambda_{\epsilon_{vt}}\Lambda_{\epsilon_{tv}} = \Lambda_{\epsilon_{vp}}\Lambda_{\epsilon_{pv}} = \Lambda_{\epsilon_1} \tag{II.18}$$

$$\Lambda_{\epsilon_{vt}}\Lambda_{\epsilon_{tw}} = \Lambda_{\epsilon_{vq}}\Lambda_{\epsilon_{qw}} = \Lambda_{\epsilon_2} \tag{II.19}$$

Ces relations définissent des conditions de traitement de certains couplages. Sous ces conditions et en utilisant la condition portant sur l'opérateur $\mathcal{C}_2$ (*ie :* $\mathcal{C}_2 \equiv (R/C_v)\mathcal{I}$), l'équation de structure discrétisée prend la forme :

$$\left\{-\frac{1}{\overline{c}_s^2}\Lambda_t^4 + \Lambda_t^2\left(\Lambda_{\epsilon_1}\nabla^2 + \frac{1}{\overline{H}^2}\Lambda_{\epsilon_{wq}}\Lambda_{\epsilon_{qw}}\mathcal{L}_v\right) + \Lambda_{\epsilon_2}\Lambda_{\epsilon_{wq}}\Lambda_{\epsilon_{qw}}\overline{N}^2\nabla^2 + \frac{\overline{c}_s^4}{\overline{H}^2}\Lambda_{\epsilon_{wq}}\mathcal{L}_v\nabla^2(\Lambda_{\epsilon_{qw}}\Lambda_{\epsilon_1} - \Lambda_{\epsilon_{qv}}\Lambda_{\epsilon_2})\right\}\tilde{\partial}w = 0 \tag{II.20}$$

Il est facilement vérifiable que la condition suivante soit toujours respectée :

$$\Lambda_{\epsilon_{qw}}\Lambda_{\epsilon_1} = \Lambda_{\epsilon_{qv}}\Lambda_{\epsilon_2} \tag{II.21}$$

Ainsi, la relation de dispersion discrète est proche du cas continu :

$$\left\{\frac{1}{\overline{c}_s^2}\Lambda_t^4 + \Lambda_t^2\left(\Lambda_{\epsilon_1}k^2 + \Lambda_{\epsilon_{wq}}\Lambda_{\epsilon_{qw}}\frac{\nu^2+1/4}{\overline{H}^2}\right) + \Lambda_{\epsilon_2}\Lambda_{\epsilon_{wq}}\Lambda_{\epsilon_{qw}}\overline{N}^2k^2\right\}w = 0 \tag{II.22}$$

Grâce à cette dernière équation, il est possible d'écrire l'ensemble des conditions de stabilité CFL :

- Les ondes de gravité sont celles qui apparaissent lorsque $\overline{c}_s$ est infiniment grand. Pour rappel, les nombres de Courant dépendent également de la direction de la propagation. Dans le cas où le nombre d'ondes verticales est négligeable devant le nombre d'ondes horizontales (*ie :* $(\ell^2+1/4)/\overline{H}^2 \ll k^2$), alors :

  $$\Delta\tau^2\Lambda_{\epsilon_1}\Lambda_t^2 + C_{g,x}^2\Lambda_{\epsilon_2}\Lambda_{\epsilon_{wq}}\Lambda_{\epsilon_{qw}} = 0 \tag{II.23}$$

  avec $C_{g,x} = \Delta\tau\overline{N}$ le nombre de Courant associé à la propagation horizontale des ondes de gravité.
  De même, en faisant l'hypothèse inverse sur les nombres d'ondes (*ie :* $k^2 \ll (\ell^2+1/4)/\overline{H}^2$), alors :

  $$\Delta\tau^2\Lambda_t^2 + C_{g,z}^2\Lambda_{\epsilon_2} = 0 \tag{II.24}$$

  avec $C_{g,z} = 2\Delta\tau\overline{HN}k$ le nombre de Courant associé à la propagation « oblique » des ondes de gravité.

- Les ondes acoustiques sont celles définies lorsque $\overline{N}$ est nulle. De plus, à supposer que $k^2 \ll (\ell^2 + 1/4)/\overline{H}^2$, alors :

$$\Delta\tau^2\Lambda_t^2 + C_{a,z}^2\Lambda_{\epsilon_{wq}}\Lambda_{\epsilon_{qw}} = 0 \tag{II.25}$$

avec $C_{a,z} = \Delta\tau\frac{\overline{c}_s}{\overline{H}}\sqrt{l^2 + \frac{1}{4}}$ le nombre de Courant associé à la propagation verticale des ondes acoustiques.
Enfin, la structure des ondes acoustiques horizontales se déduit en faisant l'hypothèse que $(\ell^2 + 1/4)/\overline{H}^2 \ll k^2$) :

$$\Delta\tau^2\Lambda_t^2 + C_*^2\Lambda_{\epsilon_1} = 0 \tag{II.26}$$

avec $C_* = \Delta\tau\overline{c}_s k$ le nombre de Courant associé à la propagation horizontale des ondes acoustiques

Le traitement implicite imposé par les schémas HEVI doit permettre de s'affranchir de la contrainte sur la stabilité de la propagation verticale des ondes acoustiques par un traitement implicite. Ainsi, l'équation de dispersion pour cette onde et pour cette direction (II.25) se réduit à :

$$(\lambda - 1)^2 + C_{a_z}^2\left(\lambda + \frac{\epsilon_{wq} + \epsilon_{qw} - \epsilon_{qw}\epsilon_{wq}}{2\epsilon_{qw}\epsilon_{wq}}\right)^2 = C_{a_z}^2\left(\frac{\epsilon_{wq} - \epsilon_{qw}}{2\epsilon_{qw}\epsilon_{wq}}\right)^2 \tag{II.27}$$

Avec un traitement ϵ_{wq} et ϵ_{qw} équivalent, il est aisé de montrer que les solutions de l'équation sont de module 1 si, et seulement si, ces coefficients sont égaux à 1/2 (traitement trapézoïdal). Ce type de traitement permet donc de ne pas amortir la propagation verticale des ondes acoustiques.

Pour les cas de l'horizontale, où un traitement explicite est appliqué par la contrainte HEVI, la relation de dispersion sur la propagation horizontale des ondes acoustiques (II.26) se résume à :

$$\left(\lambda + \frac{C_*^2\epsilon_1 - 2}{2}\right)^2 + C_*^2\frac{4 - C_*^2\epsilon_1^2}{4} = 0 \tag{II.28}$$

Afin d'avoir une solution de module 1, il est nécessaire que le second membre soit positif, ce qui impose une première condition sur la stabilité :

$$C_*\epsilon_1 \leq 2 \tag{II.29}$$

Sous cette condition, le calcul de la norme de λ nous donne :

$$\Gamma^2 = 1 + (1 - \epsilon_1)C_*^2 \tag{II.30}$$

Ainsi, si le vent est traité de manière purement implicite ($\epsilon_1 = 1$) dans les équations du bilan d'énergie (II.13) et de la continuité (II.14), la condition CFL sur les modes acoustiques horizontaux est réduite à $C_* \leq 2$. De plus, $\Lambda_{\epsilon_1} = \Lambda_{\epsilon_2}$, et les relations de dispersion discrètes sur les ondes de gravité imposent des conditions de stabilité similaires à celles-ci. Or, comme les ondes acoustiques ont une vitesse de propagation plus élevée, il suffit donc de respecter la condition CFL pour ces ondes afin d'avoir un pas de temps assurant la stabilité du schéma.

Le traitement ainsi imposé par, d'un côté, des schémas d'Euler implicites et explicites, et de l'autre, la contrainte HEVI, décrit le schéma forward-backward étudié par Messinger (1977) [47]. Il est important de noter que ce traitement implicite n'augmente en rien le nombre de calculs à effectuer pour résoudre l'algorithme. En effet, comme le vent horizontal est évalué de manière explicite, alors, une simple substitution de ce terme permet ce traitement implicite. Notons que la condition de stabilité $\epsilon_1 = 1$ peut aussi faire référence à un autre traitement. En effet, il est nécessaire que le vent horizontal soit calculé implicitement dans les équations (II.13) et (II.14) et que l'équation du mouvement (II.11) soit résolue de manière explicite, ou à l'inverse, que les termes horizontaux des équations (II.13) et (II.14) soient explicites et que l'on ait une évaluation implicite des équations du mouvement. Là encore, ce schéma n'est pas plus cher numériquement, car il revient au traitement précédent en changeant l'ordre des calculs. Ce traitement possède exactement les mêmes caractéristiques que le précédent, c'est pourquoi, il n'est présenté ici que la version forward-backward originale.

Comme les schémas HEVI ne sont pas soumis à la condition CFL portant sur la propagation verticale des ondes acoustiques, nous devons vérifier à chaque fois cette propriété. Pour cela, nous utilisons le paramètre r, introduit par Lock *et al.* (2014) [42], défini par (I.27) et variant dans l'intervalle $[10^{-2}, 10^3]$. Ce paramètre trouve sa traduction pour le nombre d'ondes verticales ℓ par l'équation suivante :

$$\ell = \sqrt{(rk\overline{H})^2 - \frac{1}{4}} \tag{II.31}$$

Par ce paramètre, il est donc possible de faire varier le nombre de Courant C_{a_z} en fonction de C_* afin de vérifier l'inconditionnelle stabilité des schémas étudiés ici en fonction de la propagation verticale des ondes acoustiques.

La présence des termes liés à l'orographie forme une différence importante entre le système linéaire $\mathscr{L}$ (utilisé pour les analyses) et celui employé de manière plus opérationnelle $\mathscr{L}^t$. En effet, du fait que la coordonnée masse soit une coordonnée épousant la forme du relief, Kasahara (1974)[32] montre que certains termes apparaissent pour compenser la variation de la coordonnée cartésienne initiale. Ces termes se retrouvent à la fois dans la divergence totale des équations de la température (II.6) et de la pression (II.7), ainsi que sur la dérivée horizontale de la pression dans les équations du mouvement horizontal (II.4). Dans cette partie, ces termes sont évalués explicitement.

Effet d'une divergence damping sur les solutions du système linéaire

Une méthode classique pour augmenter la stabilité de la méthode propose l'ajout d'un terme diffusif dans les équations du mouvement. Gassmann & Herzog (2007) [24] étudient l'effet de mettre cette *divergence damping* sur l'une ou l'autre des équations de mouvement. Leurs conclusions confortent le choix initial réalisé par Skamarock & Klemp (1992) [60] à savoir l'ajout d'une divergence 3D dans les trois équations du mouvement (II.11)-(II.12). Afin de déterminer l'impact de ces termes sur la structure des modes, il suffit d'étudier l'équation de structure dans l'espace des modes propres du système. Sur le système linéaire $\mathscr{L}$, ces filtres se présentent ainsi :

$$
\begin{aligned}
\partial_t u + R\overline{\mathcal{G}}\nabla T + R\overline{T}(\mathrm{I}-\overline{\mathcal{G}})\nabla q + \frac{R\overline{T}}{\overline{\pi_s}}\nabla\pi_s - \gamma_h \nabla D_3 &= 0 \\
\partial_t w - g(\tilde{\partial}+\mathrm{I})q + \frac{\gamma_v}{\overline{\overline{H}}}\tilde{\partial}D_3 &= 0 \\
\partial_t T + \frac{R\overline{T}}{C_v}\left(\nabla u - \frac{1}{\overline{\overline{H}}}\tilde{\partial}w\right) &= 0 \\
\partial_t q - \overline{\mathcal{S}}\nabla u + \frac{C_p}{C_v}\left(\nabla u - \frac{1}{\overline{\overline{H}}}\tilde{\partial}w\right) &= 0 \\
\partial_t \pi_s + \overline{\pi_s}\overline{\mathcal{N}}\nabla u &= 0
\end{aligned}
$$

où $D_3 = \nabla u - \tilde{\partial}w/\overline{H}$ et γ_h et γ_v sont les coefficients de la divergence damping.

La relation de dispersion définissant les propriétés des ondes décrites par ce système est modifiée par rapport au système sans filtre :

$$
\begin{aligned}
\omega^4 + \hat{\imath}\left(\gamma_h k^2 + \frac{\gamma_v}{\overline{\overline{H}}}\left(\hat{\imath}\nu - \frac{1}{2}\right)^2\right)\omega^3 - \overline{c}_s^2\left(k^2 + \frac{\nu^2+1/4}{\overline{H}^2}\right)\omega^2 & \\
+ \hat{\imath}\frac{g}{\overline{\overline{H}}}(\gamma_h - \gamma_v)\left(\hat{\imath}\nu - \frac{1}{2}\right)k^2\omega + \overline{N}^2\overline{c}_s^2 k^2 = 0 & \qquad \text{(II.32)}
\end{aligned}
$$

L'apparition des termes en ω d'ordre impair confirme que les ondes rapides du système sont bien modifiées par la présence de ce nouveau terme. Les travaux de Gassmann & Herzog (2007) [24] montrent que, pour des valeurs identiques de γ_h et γ_v, alors la fréquence des ondes les plus rapides est diminuée, et de plus, le signal est amorti. En revanche, pour le cas où γ_v est nul, l'apparition du terme d'ordre 1 modifie sensiblement la fréquence des ondes les plus lentes (et donc nécessairement les ondes de gravité), sans pour autant modifier leurs amplitudes.

Une dernière possibilité, utilisée par Klemp *et al.* (2007) [34], est d'avoir comme terme diffusif uniquement $D_3 = \nabla u$, et de ne l'utiliser que dans les équations du mouvement horizontal. Dans ce cas, l'équation de structure est particulièrement proche du système initial :

$$
\omega^4 + \hat{\imath}\gamma_h k^2\omega^3 - \overline{c}_s^2\left(k^2 + \frac{\nu^2+1/4}{\overline{H}^2}\right)\omega^2 + \overline{N}^2\overline{c}_s^2 k^2 = 0 \qquad \text{(II.33)}
$$

Le comportement des quatre ondes critiques solutions de cette dernière équation est résumé dans le Tableau (II.1).

Il est à noter que dans le cas des ondes de gravité, ou des ondes acoustiques verticales, nous n'avons aucune modification de leurs structures avec l'ajout du coefficient γ_h. Ainsi, le terme de diffusion ajouté ne déforme que les ondes acoustiques se propageant dans la direction horizontale. Pour ces dernières, la fréquence est définie par :

$$
\omega^2 = \overline{c}_s^2 k^2\left[\left(1 - \left(\frac{\gamma_h k}{2\overline{c}_s}\right)^2\right) - \hat{\imath}\frac{\gamma_h k}{2\overline{c}_s}\right]
$$

Onde	Conditions		Équation
a) acoustique verticale	$k=0$	$\overline{N}=0$	$\omega^2-\left(\frac{\overline{c}_s}{\overline{H}}\right)^2\left(\ell^2+\frac{1}{4}\right)=0$
b) gravité horizontale	$\ell=0$	$\frac{1}{4\overline{H}^2}\ll k^2$; $\overline{c}_s\to\infty$	$\omega^2-\overline{N}^2=0$
c) gravité « oblique »	$\ell=0$	$k^2\ll\frac{1}{4\overline{H}^2}$; $\overline{c}_s\to\infty$	$\omega^2-4\overline{H}^2k^2\overline{N}^2=0$
d) acoustique horizontale	$\ell=0$	$\frac{1}{4\overline{H}^2}\ll k^2$; $\overline{N}=0$	$\omega^2+\hat{\imath}\gamma_h k^2\omega-\overline{c}_s^2k^2=0$

TABLE II.1 – *Impact de la divergence damping proposée par Klemp et al. (2007) [34] sur la structure des ondes rapides du système linéarisé $\mathscr{L}$.*

Comme pour le cas du filtre initialement proposé par Skamarock & Klemp (1992) [60], le terme diffusif a pour effet, d'une part, de ralentir la vitesse de l'onde acoustique se propageant sur horizontale, et d'autre part, d'amortir son signal. À la différence de l'autre terme diffusif, celui-ci n'impacte pas les ondes acoustiques verticales. Comme le traitement implicite de la verticale assure l'inconditionnelle stabilité de cette direction pour les ondes, il semble donc plus important de déformer ces solutions le moins possible. Ainsi, ce dernier filtre semble être le meilleur candidat pour des schémas HEVI.

La partie définissant le schéma d'intégration des termes responsables de la propagation des ondes rapides est terminée. Pour appliquer les méthodes au pas de temps fractionné, il est maintenant nécessaire de savoir comment intégrer le reste du modèle.

3 Gestion temporelle des termes lents

Comme nous l'avons précédemment indiqué, ces termes lents apparaissent comme des sources durant les sous-pas de temps. La question est donc de savoir quelle est la forme de ces sources et comment manipuler les mises à jour afin d'intégrer les processus lents de manière consistante (et même avec un ordre de précision plus élevé). Pour présenter la réponse à ces questions, une première illustration du principe s'impose pour un schéma utilisant plusieurs niveaux temporels comme le saute-mouton. En utilisant la somme télescopique suivante :

$$X^{t+\Delta t}-X^{t-\Delta t}=\sum_{\tau=t-\Delta t}^{t+\Delta t-\Delta\tau}X^{\tau+\Delta\tau}-X^{\tau} \tag{II.34}$$

et en omettant ici d'écrire les termes rapides (qui ont déjà été étudiés) l'utilisation de la relation ci-avant permet d'écrire le système discrétisé temporellement :

$$\sum_{\tau=t-\Delta t}^{t+\Delta t-\Delta\tau}\frac{X^{\tau-\Delta\tau}-X^{\tau}}{\Delta\tau}=2MS(X^t) \tag{II.35}$$

Ainsi, pour que cette dernière équation soit en raccord avec (II.1) il suffit d'imposer, pour chaque itération sur les sous-pas de temps :

$$\frac{X^{\tau+\Delta\tau}-X^{\tau}}{\Delta\tau}=S(X^t) \tag{II.36}$$

Cet exemple illustre que, plus généralement, si un schéma avec plusieurs niveaux temporels Φ est utilisé pour l'évaluation des termes sources, de sorte que :

$$\frac{X^{t+\Delta t} - X^{t-\delta\Delta t}}{\Delta t} = \Phi(S(X^t), S(X^{t-\Delta t}), \cdots)$$

alors son traitement durant les sous pas de temps est donné par :

$$X^+ = X^0 + \frac{\Delta t}{M}\Phi(S(X^t), S(X^{t-\Delta t}), \cdots) \tag{II.37}$$

où X^0 et X^+ l'état du système à l'instant courant, et à l'itération suivante (*ie :* à $\tau + \Delta\tau$), et $S(X^t)$, $S(X^{t-\Delta t})$,... représentent le nombre de termes nécessaires au schéma. La clé δ définit l'état à partir duquel l'intégration est réalisée.

Ainsi, la forme de ces termes sources modélisant les processus lents sont des fractions de schémas. Ceux-ci peuvent utiliser plusieurs pas de temps (comme les méthodes saute-mouton, Kurihara, Adams-Bathford...), ou se réaliser en plusieurs étapes (Runge-Kutta). Initialement, les premières méthodes fractionnées HEVI utilisaient un saute-mouton pour les processus advectifs (Klemp & Wilhelmson (1978), Satoh (2002), Gassmann & Herzog (2007), Klemp *et al.* (2007) [24, 35, 34, 56]). D'autres méthodes plus précises ont été élaborées, notamment par Durran (1991) [19], mais toujours avec des méthodes aux multi-pas de temps.

Les schémas que nous avons retenus sont le classique saute-mouton ($\delta = 1$), et le schéma de Kurihara (K(M)-Split) qui est un schéma de prédiction-correction, dont le schéma prédictif est le saute-mouton, et dont le schéma correctif est un schéma trapézoïdal explicite ($\delta = 0$) utilisant l'état intermédiaire pour l'évaluation de la dynamique. Ce schéma est revisité dans ces travaux car il a intuitivement de meilleures propriétés sur la stabilité que le schéma saute-mouton. En effet, le nombre de Courant maximal qui permet au schéma saute-mouton d'être stable est 1, et qu'il est de $\sqrt{2}$ pour le Kurihara. Alors que ce dernier nécessite un tiers de calculs en plus (du fait de l'itération corrective supplémentaire), il permet une augmentation du nombre de Courant advectif de plus de 42%, ce qui permet, *a priori*, de faire des économies en ayant moins de mises à jour de la partie lente à effectuer. Mais, les premières analyses de stabilité réalisées par Skamarock & Klemp (1992) [60] montrent que le comportement de ce schéma devient très instable en présence d'advection. Nous pensons qu'il est possible de corriger ces instabilités, et c'est pourquoi nous allons poursuivre les études pour ce candidat trop rapidement disqualifié. Ainsi, les schémas d'intégration utilisés se résument :

$$\Phi(X_1, X_2) = \frac{1}{2}(X_1 + X_2) \tag{II.38}$$

Avec ces notations, nous pouvons écrire le saute-mouton comme $\Phi(S(X^t), S(X^t))$ et prenant soin de partir de l'état $X^{t-\Delta t}$ et de faire $2M$ itérations (de $t - \Delta t$ à $t + \Delta t$). Si nous appliquons le schéma de K(M)-Split, alors il suffit d'utiliser le terme évalué par ce premier schéma (noté $X^{(1)}$), de mettre à jour les termes sources en calculant $\Phi(S(X^t), S(X^{(1)}))$ et de réaliser le second schéma sur M itérations (de t à $t + \Delta t$). La FigII.1 illustre l'itération supplémentaire du K(M)-Split par rapport au saute-mouton.

Il faut noter que le schéma saute-mouton est régulièrement utilisé avec un filtre temporel. Les méthodes utilisant plus de deux pas temps (comme les méthodes saute-mouton et Kurihara) génèrent des solutions non-physiques lors de la résolution numérique. Les solutions supplémentaires (appelées *modes computationnels*) génèrent du bruit qui nuit à la précision du schéma, particulièrement en présence de discontinuités spatiales. Afin d'amortir ces nuisances, il est possible d'utiliser le filtre temporel de RAW (Williams (2009, 2011) [71, 72]. L'idée est d'appliquer une diffusion temporelle au schéma saute-mouton pour déterminer $\tilde{X}^+$ (la variable X^+ filtrée une première fois) et $\tilde{\tilde{X}}^0$ (la variable X^0 filtrée une seconde fois) telle que :

$$\tilde{X}^+ = X^+ + \frac{\nu(\alpha-1)}{2}(X^+ - 2\tilde{X}^0 + \tilde{\tilde{X}}^-)$$
$$\tilde{\tilde{X}}^0 = \tilde{X}^0 + \frac{\nu\alpha}{2}(X^+ - 2\tilde{X}^0 + \tilde{\tilde{X}}^-)$$

où, traditionnellement, $\nu = 0,1$ et $\alpha = 0,53$. Dans le cas où $\nu = 0$, aucun filtrage n'est appliqué. Dans le cas où $\alpha = 1$, alors il n'y a aucune modification sur la variable $X^{t+\Delta t}$ et les effets de la matrice sont semblables au filtre d'Asselin (1972) [4]. Le principal effet de ce filtre est d'amortir les modes, et plus spécifiquement le mode computationnel. Néanmoins, alors que le schéma saute-mouton est d'ordre 2, l'application de ce filtre détériore l'ordre de la précision. De plus, la stabilité est sensiblement plus faible. À la fois pour ne pas avoir à amortir le mode physique, mais aussi pour restaurer l'ordre de précision, Willams propose de fixer le coefficient α proche de $1/2$. En effet, pour des valeurs comprises entre $1/2$ et 1, plus α est petit, moins le mode physique est impacté, plus la stabilité augmente, et plus le mode computationnel est amorti. Mais, pour la valeur critique de $\alpha = 1/2$, le mode physique a toujours une amplitude strictement supérieure à 1, ce qui rend le schéma inconditionnellement instable, et donc, inutilisable.

Une alternative à ces schémas avec plusieurs pas de temps est la méthode de Runge-Kutta-3 utilisée par Wicker & Skamarock (1998, 2002) [69, 70]. L'écriture d'un tel schéma nécessite la définition d'états intermédiaires $X^{(1)}$ et $X^{(2)}$ tels que :

$$X^{(1)} = X^t + \frac{\Delta t}{3}S(X^t) \tag{II.39}$$
$$X^{(2)} = X^t + \frac{\Delta t}{2}S(X^{(1)}) \tag{II.40}$$
$$X^{t+\Delta t} = X^t + \Delta t S(X^{(2)}) \tag{II.41}$$

Comme dans le cas du K(M)-Split, tous les schémas utilisant plusieurs étapes nécessitent une ré-initialisation après chacune des étapes pour recommencer l'algorithme au temps t. Enfin, il faut noter que, dans ce cas, il est nécessaire que M soit un multiple de 6. De manière générale pour les schémas Rugne-Kutta, il est nécessaire que M soit un multiple de tous les dénominateurs divisant le pas de temps Δt. De plus, rappelons que, pour le cas d'une équation d'advection simple, ce schéma est stable pour un nombre de Courant maximal valant $\sqrt{3}$. Ainsi, si le pas de temps est découpé en un grand nombre de sous-pas de temps, il semble que ce schéma soit plus économe. Par exemple, si $M = 6$, il est nécessaire que le schéma saute-mouton fasse $2M = 12$ itérations pour calculer l'état $X^{t+\Delta t}$ à partir de X^t, alors que le schéma Runge-Kutta ne nécessite que 11 itérations.

Maintenant que nous savons comment traiter séparément l'ensemble des termes du système, il ne reste plus qu'à montrer comment intégrer ces termes sur une itération pour être capable d'appliquer cette méthode.

Intégration du système sur le sous-pas de temps

À partir du moment où tous les paramètres sont définis, il est possible de présenter la méthode de résolution de cette famille de schémas. L'opérateur $\mathscr{L}^t$ est partitionné en trois parties : $\mathscr{L}_e^t$ comprenant les équations du mouvement horizontal (II.4), $\mathscr{L}_a^t$ les termes horizontaux des divergences totales dans les équations d'énergie (II.6) et de continuité (II.7), ainsi que le terme d'intégration $\mathcal{S}^t$ (du fait que ce terme provient de la linéarisation de $\dot{\pi}/\pi$), et $\mathscr{L}_i^t$ qui contient le reste du modèle qui est traité implicitement. Ainsi, l'algorithme de résolution des méthodes du pas de temps fractionné HEVI s'écrit :

$$X^+ - \frac{\Delta\tau}{2}\mathscr{L}_i^t \cdot X^+ = \Delta\tau\mathscr{L}_a^t \cdot X^+ + X^\bullet \tag{II.42}$$

avec

$$X^\bullet = X^0 + \Delta\tau\mathscr{L}_e^t \cdot X^0 + \frac{\Delta t}{M}\Phi(S(X^t), S(X^\star)) \tag{II.43}$$

Les membres de droite évalués explicitement (et $X^\star = X^t$ dans le cas saute-mouton et $X^\star = X^{(1)}$ pour l'itération corrective de K(M)-Split). Pour le cas du Runge-Kutta-3, la définition de Φ suit la définition du schéma (II.39)-(II.41).

D'après la précédente remarque, sur le fait que la substitution induite par le traitement forward-backward n'engendrait aucune inversion, le problème à inverser à chaque itération se formalise par :

$$\left(\mathrm{I} - \frac{\Delta\tau}{2}\mathscr{L}_i^t\right) \cdot X^+ = X^{\bullet\bullet}$$

avec

$$X^{\bullet\bullet} = X^\bullet + \Delta\tau\mathscr{L}_a^t \cdot X^+$$

Grâce à ces définitions, nous pouvons écrire le système à inverser :

$$\mathbf{V}^+ = \mathbf{V}^{\bullet\bullet} \tag{II.44}$$

$$w^+ - g\frac{\Delta\tau}{2}\left(\tilde{\partial} + \mathcal{I}\right)^t (\mathrm{e}^t q^+) = w^{\bullet\bullet} \tag{II.45}$$

$$T^+ - \frac{\Delta\tau}{2}\frac{g}{C_v}\tilde{\partial}^t w^+ = T^{\bullet\bullet} \tag{II.46}$$

$$q^+ - \frac{\Delta\tau}{2H^t}\frac{C_p}{C_v}\tilde{\partial}^t w^+ = q^{\bullet\bullet} \tag{II.47}$$

$$\pi_s^+ = \pi_s^{\bullet\bullet} \tag{II.48}$$

Pour résoudre ce système linéaire, plusieurs techniques sont envisageables. La première consiste à inverser la matrice du membre de gauche et appliquer cet inverse au membre de droite. Comme cette matrice n'a aucune propriété remarquable, cette inversion est potentiellement chère numériquement. La technique de la substitution semble plus économique. Comme nous sommes en coordonnée masse, et que nous imposons comme conditions limites supérieures des conditions élastiques (conditions portant donc sur la pression $q_T = 0$), alors, il suffit, pour résoudre ce système, de substituer w^+ dans l'équation (II.47), d'où l'équation d'Helmholtz discrétisée :

$$q^+ - \left(\frac{\Delta\tau}{2}\right)^2 \left(\frac{c^t}{H^t}\right)^2 \mathcal{L}_v^t(\mathrm{e}^t q^+) = q^{\bullet\bullet\bullet} \tag{II.49}$$

avec

$$q^{\bullet\bullet\bullet} = q^{\bullet\bullet} + \frac{\Delta\tau}{2H^t}\frac{C_p}{C_v}\tilde{\partial}^t w^{\bullet\bullet} \tag{II.50}$$

et $c^t = \sqrt{RT^t C_p/C_v}$ la vitesse du son à la température T^t et $\mathcal{L}_v^t \equiv \tilde{\partial}^t \cdot \left(\tilde{\partial} + \mathcal{I}\right)^t$.

Lors de la résolution numérique de cette équation, l'opérateur $\mathcal{L}_v$ est évalué par des différences finies d'ordre deux. C'est pourquoi, cette application est semblable à une matrice tri-diagonale, et la résolution se fait directement par une méthode de double descente (algorithme de Thomas) pour chaque colonne du modèle.

Pour déterminer la faisabilité de ces méthodes, il ne reste qu'à étudier leur stabilité. Comme le but est de déterminer le schéma le plus apte à être utilisé dans un contexte opérationnel, et que les schémas étudiés ici sont tous, au moins, d'ordre deux, le meilleur candidat sera celui qui est le plus stable.

4 Stabilité des différents schémas

La condition de stabilité CFL $C_* \leq 2$ (définie par l'étude du système linéaire seul) n'est qu'une condition nécessaire à la stabilité globale du schéma, mais pas suffisante. Comme pour les méthodes additive-splitting, un pas de temps respectant simultanément toutes les contraintes CFL n'est pas une condition suffisante pour la stabilité globale du schéma. Celui-ci dépend, entre autre, du nombre d'itérations M. *A priori*, il est préférable d'avoir un nombre M le plus grand possible pour faire la plus grande économie de calculs. Mais, si M est grand, cela signifie que les processus évoluant lentement sont contraints à rester stationnaires, ce qui, au mieux, peut nuire à la qualité de l'intégration, et, au pire, peut engendrer des instabilités dues à la contrainte CFL portant sur le nombre de Courant advectif. Ainsi, un équilibre est à déterminer entre l'économie de calculs souhaitée sur les processus lents, et le nombre total d'itérations à effectuer pour réaliser une prévision à une échéance fixée.

Pour réaliser ces études de stabilité, nous restons focalisés sur le système linéaire sans bord $l\mathscr{L}$, que nous découpons de la même manière que précédemment, avec un écoulement constant à vitesse $\overline{U}$. Le sous-pas de temps critique $\Delta\tau$ est calculé en fonction de C_*, et la méthode la plus stable est donc celle qui permet d'avoir ce nombre de Courant le plus grand possible. D'après l'étude

sans advection, il semble que le schéma le plus performant, en terme de stabilité, soit celui dont l'apport des termes sources renforce le moins possible la condition $C_* \leq 2$.

Comme certains schémas que nous étudions ici nécessitent la connaissance de deux états $X^{t-\Delta t}$ et X^t, nous introduisons le vecteur généralisé $Y^t = {}^{\mathrm{t}}(X^{t-\Delta t}; X^t; X^{t-\Delta t})$ et nous mettons le schéma sous la forme :

$$Y^{t+\Delta t} = A \cdot Y^t$$

La stabilité du schéma se détermine en fonction des valeurs propres de la matrice d'amplification A. Pour construire la matrice d'amplification, il suffit de suivre l'algorithme défini par II.42. Soient les matrices $\mathcal{H}$, $\mathcal{E}$ et $\mathcal{P}_1$ définies par :

$$\mathcal{H} = \begin{pmatrix} \mathrm{I} - \hat{\imath} C_* (l\mathscr{L}_i/2 + l\mathscr{L}_b) & 0 & 0 \\ 0 & \mathrm{I} & 0 \\ 0 & 0 & \mathrm{I} \end{pmatrix}; \mathcal{E}_{SM} = \begin{pmatrix} \mathrm{I} + \hat{\imath} C_* (l\mathscr{L}_e + l\mathscr{L}_i/2) & \hat{\imath} C_* M_U \mathrm{I} & 0 \\ 0 & \mathrm{I} & 0 \\ 0 & 0 & \mathrm{I} \end{pmatrix}$$

$$\mathcal{P}_1 = \begin{pmatrix} 0 & \mathrm{I} & 0 \\ \mathrm{I} & 0 & 0 \\ 0 & \mathrm{I} & 0 \end{pmatrix}$$

Avec ces matrices, la matrice d'amplification du schéma saute-mouton s'écrit :

$$A_{SM} = \mathcal{P}_1 \cdot \left(\mathcal{H}^{-1} \cdot \mathcal{E}_{SM} \right)^{2M}$$

Dans le cas du schéma K(M)-Split, il suffit d'introduire deux nouvelles matrices, et d'appliquer un nouveau schéma trapézoïdal au schéma précédent. Soient $\mathcal{E}_K$ et $\mathcal{P}_2$ telles que :

$$\mathcal{E}_K = \begin{pmatrix} \mathrm{I} + \hat{\imath} C_* (l\mathscr{L}_e + l\mathscr{L}_i/2) & \hat{\imath}(C_*/2) M_U \mathrm{I} & \hat{\imath}(C_*/2) M_U \mathrm{I} \\ 0 & \mathrm{I} & 0 \\ 0 & 0 & \mathrm{I} \end{pmatrix}; \mathcal{P}_2 = \begin{pmatrix} 0 & 0 & \mathrm{I} \\ \mathrm{I} & 0 & 0 \\ 0 & 0 & \mathrm{I} \end{pmatrix}$$

Alors, la matrice d'amplification du schéma de K(M)-Split s'écrit :

$$A_K = \mathcal{P}_2 \cdot \left(\mathcal{H}^{-1} \cdot \mathcal{E}_K \right)^M \cdot A_{SM}$$

Dans ces deux schémas, dans le cas où le filtre RAW est utilisé, nous définissons une nouvelle matrice $\mathcal{F}$:

$$\mathcal{F} = \begin{pmatrix} (1 + \nu(\alpha - 1)/2)\mathrm{I} & \nu(1 - \alpha)\mathrm{I} & (\nu(\alpha - 1)/2)\mathrm{I} \\ (\nu\alpha/2)\mathrm{I} & (1 - \nu\alpha)\mathrm{I} & (\nu\alpha/2)\mathrm{I} \\ 0 & 0 & \mathrm{I} \end{pmatrix}$$

Cette matrice s'applique aux variables après le schéma saute-mouton. Ainsi, la matrice de substitution $\mathcal{P}_1$ est remplacée par $\tilde{\mathcal{P}}_1 = \mathcal{P}_1 \cdot \mathcal{F}$.

Reste à monter la matrice d'amplification de Runge-Kutta-3. En suivant le schéma défini ci-avant, et en redéfinissant les opérateurs de passage $\mathcal{P}_1$ et $\mathcal{P}_2$, ainsi que le vecteur généralisé initial $Y^t = {}^{\mathrm{t}}(X^t; X^t; X^t)$, et en introduisant une nouvelle matrice $\mathcal{P}_3$:

$$\mathcal{P}_1 = \begin{pmatrix} 0 & \mathrm{I} & 0 \\ \mathrm{I} & 0 & 0 \\ 0 & 0 & \mathrm{I} \end{pmatrix}; \mathcal{P}_2 = \begin{pmatrix} 0 & 0 & \mathrm{I} \\ 0 & \mathrm{I} & 0 \\ \mathrm{I} & 0 & 0 \end{pmatrix}; \mathcal{P}_3 = \begin{pmatrix} \mathrm{I} & 0 & 0 \\ \mathrm{I} & 0 & 0 \\ \mathrm{I} & 0 & 0 \end{pmatrix}$$

ainsi que les matrices explicites :

$$\mathcal{E}_{RK1} = \begin{pmatrix} \mathrm{I} + \hat{\imath}C_*(l\mathscr{L}_e + l\mathscr{L}_i/2) & \hat{\imath}C_* M_U \mathrm{I} & 0 \\ 0 & \mathrm{I} & 0 \\ 0 & 0 & \mathrm{I} \end{pmatrix}; \mathcal{E}_{RK2} = \begin{pmatrix} \mathrm{I} + \hat{\imath}C_*(l\mathscr{L}_e + l\mathscr{L}_i/2) & 0 & \hat{\imath}C_* M_U \mathrm{I} \\ 0 & \mathrm{I} & 0 \\ 0 & 0 & \mathrm{I} \end{pmatrix}$$

alors, la matrice d'amplification A_{RK} se définit par :

$$A_{RK} = \mathcal{P}_3 \cdot \left(\mathcal{H}^{-1} \cdot \mathcal{E}_{RK2}\right)^M \cdot \mathcal{P}_2 \cdot \left(\mathcal{H}^{-1} \cdot \mathcal{E}_{RK1}\right)^{M/2} \cdot \mathcal{P}_1 \cdot \left(\mathcal{H}^{-1} \cdot \mathcal{E}_{RK1}\right)^{M/3}$$

La FigII.2 montre le coefficient d'amplification maximal Γ en fonction du nombre de Courant C_* et M_U pour plusieurs valeurs de sous-pas de temps M et pour un nombre d'ondes verticales ℓ définies par (II.31) variant de sorte que r parcourt l'intervalle $[10^{-2}, 10^3]$. Les zones blanches correspondent à une valeur de Γ inférieure ou égale à $1 + 10^{-6}$ (*ie :* que le schéma est stable), et, à l'inverse, les zones grisées sont celles où le schéma est instable. Les figures en haut montrent la stabilité sans aucun filtre, celles du milieu montrent la même étude lorsque le filtre temporel d'Asselin est appliqué, et enfin, les dernières figures correspondent aux cas où la divergence damping de Klemp (2007) [34] est ajoutée. Cette première étude décrit trois choses. La première est que, de manière générale, nous observons que la stabilité décroît en fonction de M. Cela est dû au fait que le grand pas de temps Δt croît en fonction de M. Or, comme la vitesse d'advection est constante, alors la contrainte de stabilité portant sur le nombre de Courant de l'advection horizontale devient de plus en plus forte à mesure que M augmente. Pour autant, il semble qu'il existe toujours un pas de temps qui assure la stabilité du schéma saute-mouton. La seconde est que, comme attendu, l'effet des différents filtres tendent à augmenter la stabilité du schéma. Enfin, et à la différence de l'étude réalisée par Skamarock & Klemp (1992) [60], pour laquelle, les auteurs avaient fait la même étude, mais uniquement pour le système simple de la propagation horizontale 1D des ondes acoustiques, il existe certaines zones d'instabilité supplémentaires. Néanmoins, ces instabilités n'impactent pas la stabilité globale du schéma car elles se trouvent au-delà du nombre de Courant maximum assurant la stabilité. Il faut également noter le faible impact sur la stabilité du filtre d'Asselin, qui devient vraiment important seulement pour des nombres élevés de M. Le plus fort gain de stabilité résulte à l'ajout de la divergence damping. Pour cet exemple, nous avons utilisé le même coefficient que Klemp (2007) [34], à savoir $\gamma_h = (\Delta x/10)2/\Delta\tau$. Là encore, plus la valeur de M est élevée, plus ce filtre est efficace. En effet, plus M est grand, plus il y a d'ajout de terme diffusif sur le grand pas de temps, et donc, plus les ondes rapides ont été ralenties.

Comme ce schéma est utilisé en opérationnel (WRF), il peut nous servir de référence pour comparer la stabilité avec d'autres schémas. Un autre schéma est aussi utilisé en opérationnel,

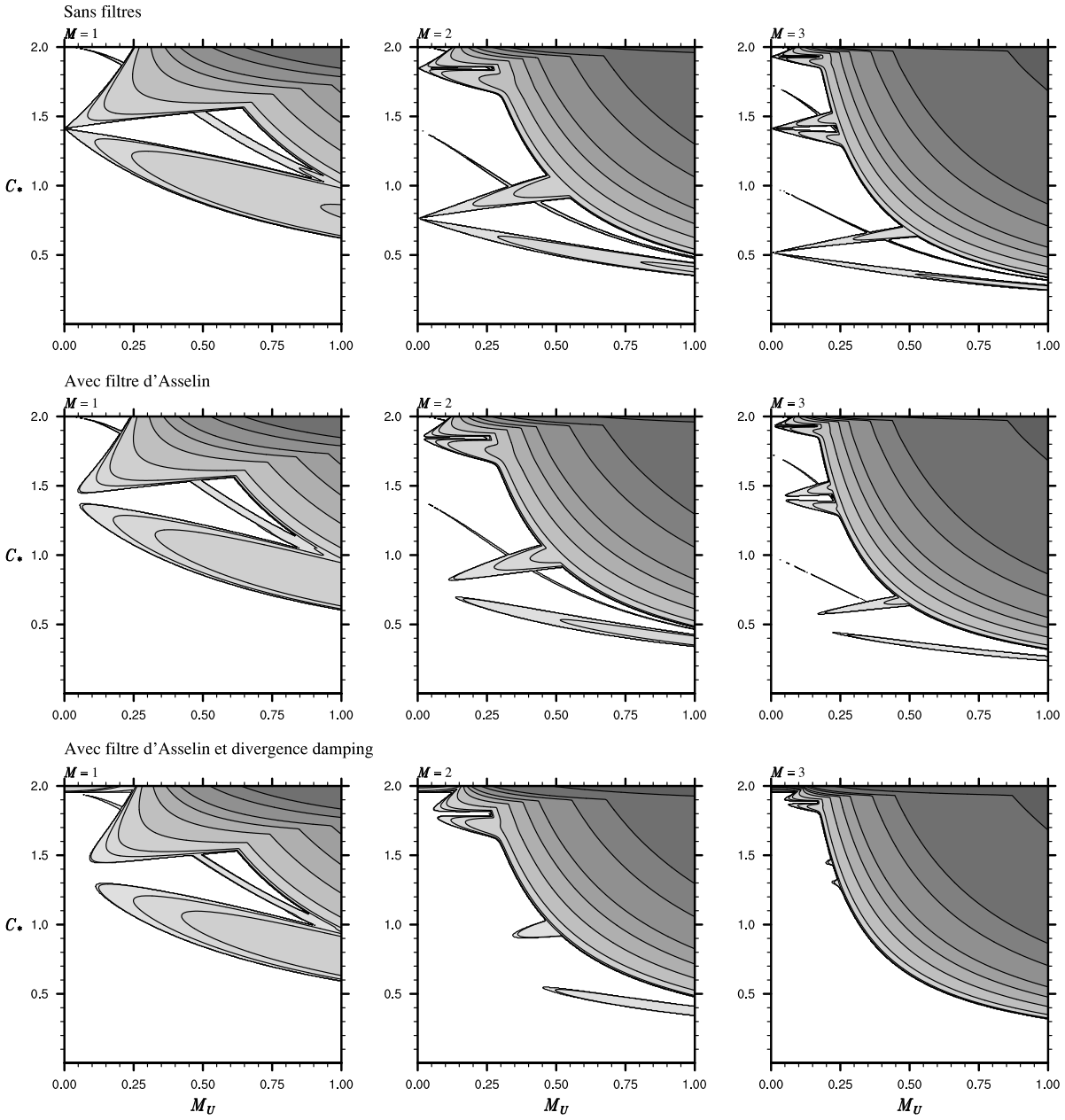


FIGURE II.2 – *Coefficient d'amplification* Γ *du schéma saute-mouton en fonction du nombre de Courant* C_* *et* M_U *pour différentes valeurs de* M*. Les zones blanches correspondent à des taux d'amplification tels que* $\Gamma < 1 + 10^{-6}$.

celui utilisant le Runge-Kutta-3, dont la stabilité est illustrée par FigII.3. Pour ce schéma, nous avons pris $M \in \{6; 12; 18\}$, car M doit être un multiple de 2 et de 3. Nous pouvons observer la très faible stabilité offerte par ce schéma en absence de diffusion. En effet, y compris pour une très faible valeur du vent, le nombre de Courant maximal chute brutalement de 2 à 1 pour $M = 6$ et même en dessous de 0,5 pour des valeurs de M supérieures à 12. Enfin, il faut noter la présence d'instabilité pour un nombre de Mach supérieur à 0,5. Ce comportement est dû principalement à

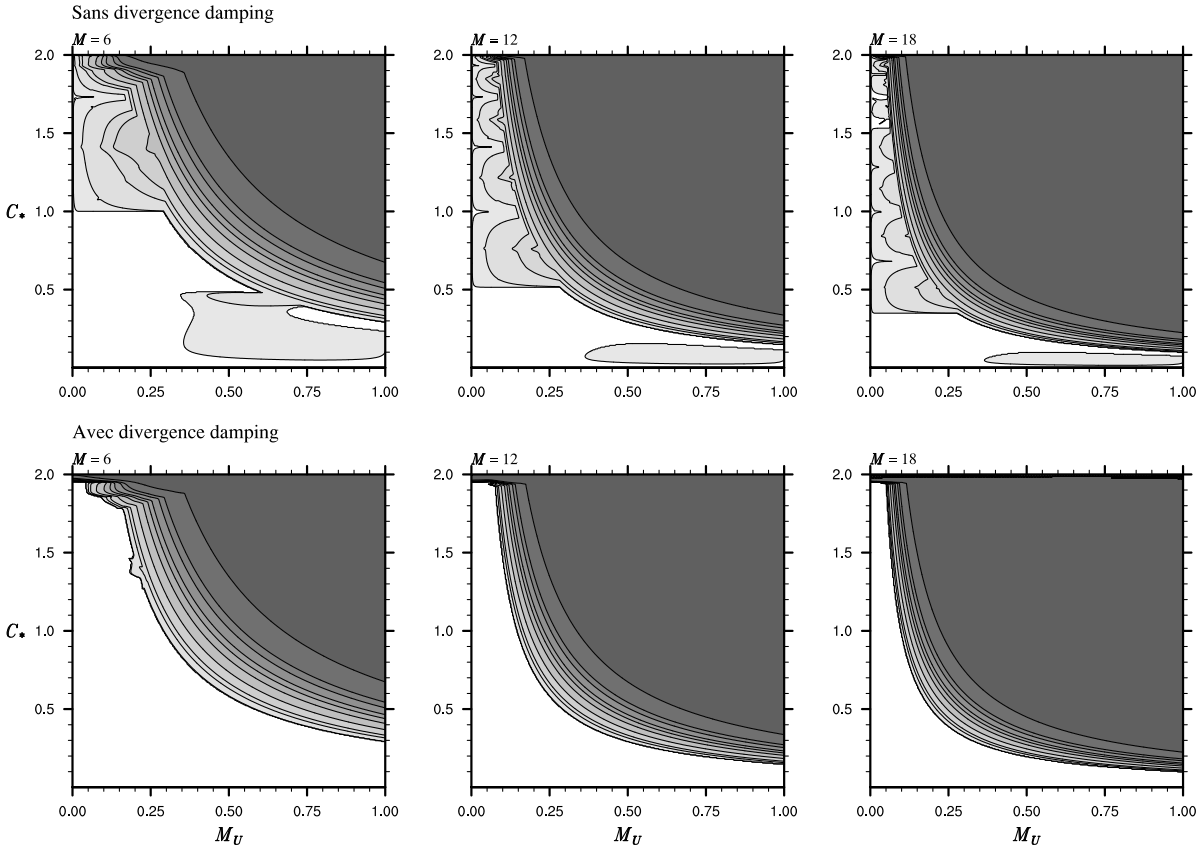


FIGURE II.3 – *Même graphique que FigII.2, pour le schéma Runge-Kutta-3.*

la nature centrée des schémas spectraux et ces instabilités sont effacées avec des schémas décentrés up-wind qui amortissent le signal comme l'expliquent à la fois Skamarock & Klemp (1992) mais aussi Wicker & Skamarock (1998, 2002) [60, 69, 70]. L'autre alternative consiste en l'utilisation de la divergence damping, qui réussit efficacement à stabiliser le schéma.

La FigII.4 montre la même étude, mais réalisée pour K(M)-Split. Nous observons, comme Skamarock & Klemp (1992) [60], que ce schéma est fortement instable en absence du filtre d'Asselin. Alors que les auteurs des analyses du schéma saute-mouton se sont arrêtés à ce résultat, il fallait poursuivre l'étude en ajoutant le filtre d'Asselin qui, d'une part, affaiblit considérablement le mode computationnel, et d'autre part ne peut qu'augmenter la stabilité de ce candidat. De plus, et à la différence du saute-mouton, cet ajout n'impacte pas l'ordre de précision du schéma. En effet, une condition suffisante pour qu'un schéma prédicteur-correcteur soit d'ordre n et que, d'une part, le schéma correctif soit d'ordre n, et que, d'autre part, le schéma prédictif soit d'un ordre au moins $n-1$. En vertu du fait que le schéma correctif soit trapézoïdal (ordre 2) et que le prédictif est d'ordre 1 (saute-mouton filtré), alors il est clair que K(M)-Split est d'ordre 2. L'ajout de ce simple artifice permet d'améliorer considérablement la stabilité de ce schéma, sans impacter la précision, et le rend le plus apte à un contexte opérationnel. Il faut néanmoins relever des zones d'instabilité autour de très grandes valeurs d'advection $M_U \leq 0{,}75$, qui sont facilement éliminées par l'ajout d'une diffusion.

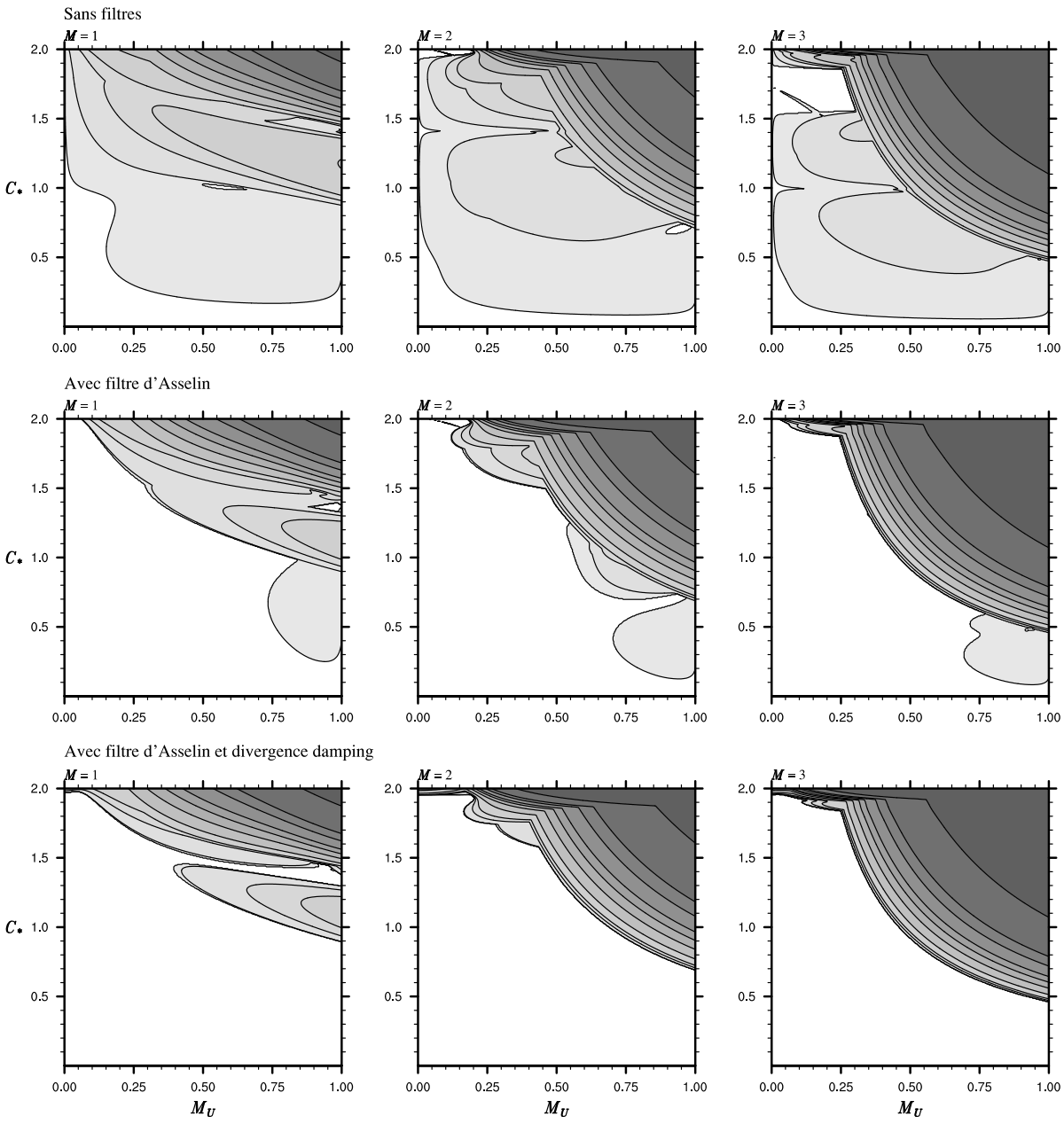


FIGURE II.4 – *Même graphique que FigII.2, mais pour le schéma K(M)-Split.*

5 Discussion

Dans ce chapitre, nous avons étudié la faisabilité des méthodes au pas de temps fractionné pour la résolution du système d'Euler pleinement compressible en coordonnée masse, avec la contrainte HEVI. Nous avons exhumé un schéma trop peu étudié dans la littérature : le K(M)-Split utilisant le filtre d'Asselin. Il permet d'une part, une économie de calculs grâce à sa meilleure stabilité y compris pour des vents importants ($M_U \leq 0{,}75$), ainsi que la précision d'ordre 2 en temps recherchée en PNT, contrairement au saute-mouton qui perd cette précision par l'introduction du filtre d'Asselin. En revanche, l'ensemble de ces méthodes semblent souffrir d'une forte contrainte sur la stabilité pour de très grandes valeurs du vent ($M_U > 0{,}75$).

Rappelons que ces méthodes sont fondées sur l'hypothèse de la grande différence entre les vitesses de propagation des processus, ce qui n'est pas toujours vérifié, y compris pour les systèmes raides de la PNT. En effet, si la vitesse des ondes rapides (comme les ondes acoustiques) évolue faiblement, ce n'est pas le cas du transport dû à l'advection qui peut varier d'un état quasiment au repos jusqu'à des jets valant plusieurs centaines de mètres par seconde (et ainsi évoluer à une vitesse proche de celle du son) qui sont présents dès 30 km d'altitude. Pour des modèles ayant des toits suffisamment hauts, il semble donc que le nombre d'itérations M à utiliser soit 1. Ceci nous amène donc à étudier les méthodes très générales Implicite-Explicite (IMEX) dont nous allons parler dans le chapitre suivant.

Chapitre III

Analyse de stabilité des méthodes RK-IMEX HEVI

Une des méthodes couramment utilisée pour résoudre des systèmes raides est l'utilisation simultanée d'un schéma implicite intégrant les termes responsables de la plus grosse contrainte de stabilité CFL, et un autre schéma traitant explicitement le reste des équations. On parle alors de méthode Implicite-Explicite (IMEX). Ceci permet d'avoir des schémas relativement stables, sans pour autant avoir une méthode trop coûteuse numériquement. De manière très générale, les méthodes IMEX nécessitent l'utilisation de plusieurs schémas dont au moins un implicite, et d'autres explicites. Les travaux présents dans la littérature ne proposent d'étudier que des méthodes IMEX utilisant seulement deux schémas. Il faut néanmoins remarquer que, *a priori*, il peut y avoir autant de schémas que de termes dans le système auquel s'applique la méthode. Ainsi, les systèmes résolus par ces méthodes s'écrivent sous la forme :

$$\partial_t X = \mathscr{M}(X) = \mathscr{E}(X) + \mathscr{I}(X) \tag{III.1}$$

où $\mathscr{E}$ représente les termes traités de manière explicite et $\mathscr{I}$ représente les termes subissant un traitement implicite.

Cette famille, très large, regroupe une grande partie des schémas opérationnels (y compris les schémas SI). Des études spécifiques de schémas IMEX respectant la contrainte HEVI ont été réalisées durant ces dernières années (Ascher *et al.* (1995,1997), Pareschi & Russo (2005), Durran & Blossey (2012) [3, 2, 20, 50]). Certains schémas, issus de ces travaux, ont été montrés comme ayant de relatives bonnes propriétés de stabilité, ou de coût, suggérant une utilisabilité efficace pour des modèles opérationnels. D'autres travaux, notamment ceux de Ullrich & Jablonowski (2012) [67] et Giraldo *et al.* (2013) [25] ont élaboré des schémas (employés aujourd'hui dans des modèles opérationnels) qui ont étaient démontrés, par Weller *et al.* (2013) [68] et Lock *et al.* (2014) [42], comme étant les schémas les plus efficaces. En revanche, ces premières études ne s'opèrent que sur des systèmes d'équations très simplifiés, qui ne prennent pas en compte certains processus déstabilisant le schéma (comme l'advection). De plus, elles se bornent à regarder l'impact de ces schémas exclusivement sur les ondes acoustiques, qui ont une faible importance météorologique.

Dès lors, plusieurs questions restent encore à poser. Quel est le comportement de la stabilité de ces schémas pour des systèmes plus complexes ? Quelle est la déformation des ondes de gravité ?

À la lumière de ces réponses, il faut encore pouvoir déterminer quel schéma reste le plus efficace.

L'objectif de ce chapitre est de prolonger les études réalisées jusqu'ici sur ces schémas, afin de déterminer le plus stable, et donc mieux adapté à la PNT. Ce chapitre se décompose en plusieurs parties. D'abord, il faut faire l'état de l'art des méthodes IMEX HEVI et sélectionner les schémas qui semblent les plus efficaces pour le système des ondes acoustiques 2D. Comme ce type d'ondes demeure aussi présent dans système linéaire $\mathscr{L}$, une stabilité relativement grande, pour le système des ondes acoustiques, est donc une condition nécessaire à une bonne stabilité du schéma appliqué pour $\mathscr{L}$. Puis, après des études sur la stabilité des candidats retenus pour le système linéaire complet (contenant les processus advectifs traités de manière explicite), nous pourrons d'une part, restreindre la liste des candidats, et d'autre part, proposer de nouveaux schémas que nous démontrerons comme étant plus stables. Dès lors, après une vérification sur l'impact de ces schémas sur les ondes de gravité, il nous sera possible de présenter le schéma qui semble le mieux à même d'être exploité dans un cadre opérationnel.

1 Présentation de la méthode

Les premiers schémas à avoir été étudiés avec ces méthodes sont ceux possédant plusieurs niveaux temporels (Ascher *et al.* (1995) [3]). Ces travaux ont été repris et augmentés par Durran & Blossey (2012) [20] qui montrent que, pour l'ensemble de ces schémas, des modes computationnels sont présents et nécessitent donc, par une manipulation artificielle telle que l'ajout de filtres, d'amortir le signal de ces modes non-physiques. Ces différents filtres détériorent la qualité de l'intégration numérique, notamment par la perte d'un ordre de précision. Pour éviter cet écueil, il semble préférable de se concentrer sur les schémas d'avance temporelle Runge-Kutta utilisant plusieurs étapes. Nous parlons alors de schéma RK-IMEX.

Méthodes avec deux tableaux de Butcher

Comme nous appliquons un schéma purement explicite pour la partie $\mathscr{E}$ et un schéma implicite $\mathscr{I}$, la discrétisation temporelle de (III.1) s'écrit uniquement à l'aide de deux schémas :

$$\frac{X^+ - X^0}{\Delta t} = \sum_{j=1}^{\nu} \tilde{b}_j \mathscr{E}(t + \tilde{c}_j \Delta t, X^{(j)}) + \sum_{j=1}^{\nu} b_j \mathscr{I}(t + c_j \Delta t, X^{(j)}) \tag{III.2}$$

$$\frac{X^{(j)} - X^0}{\Delta t} = \sum_{i=1}^{j-1} \tilde{a}_{ji} \mathscr{E}(t + \tilde{c}_i \Delta t, X^{(i)}) + \sum_{i=1}^{j} a_{ji} \mathscr{I}(t + c_i \Delta t, X^{(j)}) \tag{III.3}$$

avec ν, le nombre total d'étapes du schéma RK-IMEX, i et j sont des entiers tels que $1 \leq i \leq j \leq \nu$ et $X^{(j)}$ est la valeur du système à la j-ième étape. Les coefficients surmontés d'un tilde font référence au schéma explicite, alors que ceux sans tilde se réfèrent au schéma implicite. Les vecteurs $\tilde{b}$ et b sont les vecteurs de poids lors de l'étape finale et les coefficients $\tilde{c}$ et c marquent l'avancement temporel. Enfin, avec les matrices $\tilde{\mathcal{A}} = (\tilde{a}_{ij})$ et $\mathcal{A} = (a_{ij})$, l'ensemble de ces coefficients définissent les deux tableaux de Butcher $\{\tilde{c}, \tilde{\mathcal{A}}, \tilde{b}\}$ et $\{c, \mathcal{A}, b\}$ résumant ce schéma RK-IMEX :

$$
\begin{array}{c|ccccc}
\tilde{c}_1 & 0 & & & & \\
\tilde{c}_2 & \tilde{a}_{21} & 0 & & & \\
\vdots & \vdots & & \ddots & & \\
\tilde{c}_j & \tilde{a}_{j1} & \cdots & \tilde{a}_{jl} & \cdots & 0 \\
\vdots & \vdots & & & & \ddots \\
\tilde{c}_\nu & \tilde{a}_{\nu 1} & \cdots & \tilde{a}_{\nu l} & \cdots & 0 \\
\hline
 & \tilde{b}_1 & \cdots & \tilde{b}_l & \cdots & \tilde{b}_\nu
\end{array}
\qquad
\begin{array}{c|ccccc}
c_1 & a_{11} & & & & \\
c_2 & a_{21} & a_{22} & & & \\
\vdots & \vdots & & \ddots & & \\
c_j & a_{j1} & \cdots & a_{jl} & \cdots & a_{jj} \\
\vdots & \vdots & & & & \ddots \\
c_\nu & a_{\nu 1} & \cdots & a_{\nu l} & \cdots & a_{\nu\nu} \\
\hline
 & b_1 & \cdots & b_l & \cdots & b_\nu
\end{array}
$$

Pour assurer que ces schémas définissent bien une méthode RK-IMEX, il suffit que la matrice $\tilde{\mathcal{A}}$ soit triangulaire strictement inférieure. C'est pourquoi, dans l'équation (III.3), la somme sur l'opérateur $\mathscr{E}$ va seulement jusqu'à $j-1$. Dans le cas de systèmes physiques, qui ne dépendent pas explicitement du temps, il est d'usage courant de définir les coefficients $\tilde{c}_j$ et c_j tels que :

$$\tilde{c}_j = \sum_{i=1}^{j-1} \tilde{a}_{ji} \qquad c_j = \sum_{i=1}^{j} a_{ji}$$

Les premiers travaux étudiant la faisabilité de ces méthodes sont dus à Ascher *et al.* (1997) [2], mais surtout à Pareschi & Russo (2005) [50], qui mettent en évidence les conditions nécessaires et suffisantes que doivent vérifier les tableaux de Butcher afin d'assurer que le schéma RK-IMEX soit consistant, d'ordre 2, et même d'ordre 3. Leurs travaux sont à la base de la construction de nombreux schémas, et c'est pourquoi il est important de les présenter.

Soit $e = (1, \ldots, 1)$ le vecteur unité de dimension ν, alors le schéma RK-IMEX est consistant si, et seulement si :

$$\tilde{b}^{\mathrm{t}} \cdot e = b^{\mathrm{t}} \cdot e = 1$$

De plus, il est d'ordre 2 si, et seulement si, il est consistant et vérifie également :

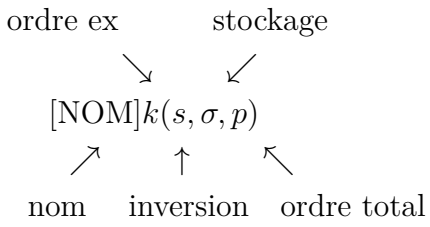
$$b^{\mathrm{t}} \cdot \mathcal{A} \cdot e = \tilde{b}^{\mathrm{t}} \cdot \tilde{\mathcal{A}} \cdot e = \frac{1}{2}$$
$$\tilde{b}^{\mathrm{t}} \cdot \mathcal{A} \cdot e = b^{\mathrm{t}} \cdot \tilde{\mathcal{A}} \cdot e = \frac{1}{2}$$

La première équation met en évidence la nécessité que chacun des schémas soit d'ordre 2 pour que le schéma RK-IMEX le soit également. La seconde équation révèle des conditions portant sur la combinaison des schémas. Enfin, le schéma est d'ordre 3 si, et seulement si, il est d'ordre 2 et vérifie :

$$\begin{aligned}
{}^{\intercal}b \cdot \mathcal{C}^2 \cdot e = {}^{\intercal}\tilde{b} \cdot \tilde{\mathcal{C}}^2 \cdot e = \frac{1}{3} \; ; \quad {}^{\intercal}b \cdot \mathcal{A} \cdot \mathcal{C} \cdot e = {}^{\intercal}\tilde{b} \cdot \tilde{\mathcal{A}} \cdot \tilde{\mathcal{C}} \cdot e = \frac{1}{6} \\
{}^{\intercal}\tilde{b} \cdot \mathcal{A} \cdot \mathcal{C} \cdot e = {}^{\intercal}b \cdot \tilde{\mathcal{A}} \cdot \mathcal{C} \cdot e = {}^{\intercal}b \cdot \mathcal{A} \cdot \tilde{\mathcal{C}} \cdot e = {}^{\intercal}\tilde{b} \cdot \tilde{\mathcal{A}} \cdot \mathcal{C} \cdot e = {}^{\intercal}\tilde{b} \cdot \mathcal{A} \cdot \tilde{\mathcal{C}} \cdot e = {}^{\intercal}b \cdot \tilde{\mathcal{A}} \cdot \tilde{\mathcal{C}} \cdot e = \frac{1}{6} \\
{}^{\intercal}\tilde{b} \cdot \mathcal{C}^2 \cdot e = {}^{\intercal}\tilde{b} \cdot \tilde{\mathcal{C}} \cdot \mathcal{C} \cdot e = {}^{\intercal}b \cdot \tilde{\mathcal{C}}^2 \cdot e = \frac{1}{3}
\end{aligned}$$

avec $\tilde{\mathcal{C}}$ et $\mathcal{C}$, les matrices diagonales de taille $\nu \times \nu$ où le coefficient de la j-ième ligne est égal au j-ième coefficient du vecteur respectivement $\tilde{c}$ et c.
Là encore, ces conditions montrent que l'utilisation de schémas d'ordre 3 est une condition nécessaire, mais pas suffisante pour garantir une précision au troisième ordre du schéma RK-IMEX.

Les conditions de Pareschi & Russo (2005) [50] ne sont pas suffisantes pour déterminer l'ensemble des coefficients des deux tableaux de Butcher. Il existe encore de nombreux degrés de liberté qui permettent ainsi l'élaboration de nombreux schémas. Ainsi, Pareschi & Russo (2005) [50] instaurent une nomenclature afin de nommer n'importe quel schéma RK-IMEX avec deux tableaux de Butcher :



Jusqu'à présent, aucun schéma RK-IMEX ayant un nombre plus important de tableaux de Butcher n'a été proposé pour intégrer les équations de la dynamique du temps. *A priori*, il est possible d'utiliser autant de tableaux qu'il y a de variables dans le modèle. Mais cette voie n'a jamais été explorée du fait qu'aucune étude sur la généralisation des travaux de Pareschi & Russo (2005) [50] n'a été réalisée. Afin d'ouvrir la voie à l'exploitation de nouveaux schémas, nous proposons d'établir les conditions nécessaires et suffisantes pour déterminer si un schéma RK-IMEX, utilisant N tableaux de Butcher, est consistant et d'ordre 2. Il semble peu utile d'étendre ces résultats aux ordres supérieurs du fait qu'à l'heure actuelle, les méthodes d'ordre 2 sont suffisantes pour la PNT.

Généralisation pour N tableaux de Butcher

Soit un système d'équations différentielles, dépendant du temps, et utilisant N termes :

$$y'(t) = \sum_{n=1}^{N} \mathcal{F}_n(t, y(t)) \tag{III.4}$$

où $\mathcal{F}_n$ correspond aux différentes contributions dynamiques, et avec $n \in [\![1; N]\!]$ représentant les différents processus. La solution exacte du système (III.4) au temps $t = t + \Delta t$ peut être approchée par $y(t)$ via l'utilisation des développements de Taylor :

$$\begin{aligned} y(t+\Delta t) &= y(t) + \Delta t\, y'(t) + \frac{\Delta t^2}{2} y''(t) + O(\Delta t^3) \\ &= y(t) + \Delta t \sum_{n=1}^{N} \mathcal{F}_n(t, y(t)) + \frac{\Delta t^2}{2} \sum_{n=1}^{N} \Bigg[\partial_t \mathcal{F}_n(t, y(t)) \\ &\qquad + \partial_y \mathcal{F}_n(t, y(t)) \cdot \left(\sum_{m=1}^{N} \mathcal{F}_m(t, y(t)) \right) \Bigg] + O(\Delta t^3) \end{aligned} \tag{III.5}$$

où, $\partial_y \mathcal{F}_n(t, y(t))$ est le Jacobien[1] de l'opérateur $\mathcal{F}_n$ appliqué au point $(t, y(t))$, et donc $\partial_y \mathcal{F}_n(t, y(t)) \cdot \mathcal{F}_m(t, y(t))$ représente de Jacobien appliqué à $\mathcal{F}_m$ au point $(t, y(t))$.

Le système (III.4) est discrétisé en N schémas Runge-Kutta avec $\nu > 1$ étapes pour chaque terme. Ainsi, le schéma intégrant le n-ième terme (représenté par $\mathcal{F}_n$) est décrit par le n-ième tableau de Butcher $\{c^{(n)}, \mathcal{A}^{(n)}, b^{(n)}\}$. Ainsi, la discrétisation en N schémas Runge-Kutta peut donc s'écrire :

$$\begin{aligned} y^{(j)} &= y^0 + \Delta t \sum_{n=1}^{N} \sum_{i=1}^{\nu} a_{ji}^{(n)} \mathcal{F}_n(t + c_i^{(n)} \Delta t, y^{(i)}) \\ y^{+} &= y^0 + \Delta t \sum_{n=1}^{N} \sum_{j=1}^{\nu} b_{j}^{(n)} \mathcal{F}_n(t + c_j^{(n)} \Delta t, y^{(j)}) \end{aligned}$$

avec $n \in [\![1; N]\!]$, $y^+ = y(t + \Delta t)$, $y^{(j)} = y(t + c_j \Delta t)$ et $y^0 = y(t)$.
En utilisant le développement de Taylor sur la fonction $\mathcal{F}_n$, il est facile d'établir que :

$$\begin{aligned} \mathcal{F}_n(t + c_j^{(n)} \Delta t, y^{(j)}) =& \mathcal{F}_n(t, y^0) + c_j^{(n)} \Delta t \partial_t f \mathcal{F}_n(t, y^0) \\ &+ \Delta t \partial_y \mathcal{F}_n(t, y^0) \cdot \left(\sum_{m=1}^{N} \sum_{i=1}^{\nu} a_{ji}^{(n)} \mathcal{F}_m(t, y^0) \right) + O(\Delta t^2) \end{aligned} \tag{III.6}$$

De là, la discrétisation finale s'écrit :

$$\begin{aligned} y^+ &= y^0 + \Delta t \sum_{n=1}^{N} \sum_{j=1}^{\nu} b_j^{(n)} \Bigg[\mathcal{F}_n(t, y^0) + c_j^{(n)} \Delta t \partial_t \mathcal{F}_n(t, y^0) + \\ &\qquad \partial_y \mathcal{F}_n(t, y^0) \cdot \left(\sum_{m=1}^{N} \sum_{i=1}^{\nu} a_{ji}^{(m)} \mathcal{F}_m(t, y^0) \right) + O(\Delta t^2) \Bigg] \\ &= y^0 + \Delta t \sum_{n=1}^{N} \left(\sum_{j=1}^{\nu} b_j^{(n)} \right) \mathcal{F}_n(t, y^0) + \Delta t^2 \sum_{n=1}^{N} \Bigg[\left(\sum_{j=1}^{\nu} b_j^{(n)} c_j^{(n)} \right) \partial_t \mathcal{F}_n(t, y^0) + \\ &\qquad \partial_y \mathcal{F}_n(t, y^0) \cdot \sum_{m=1}^{N} \left(\left[\sum_{j=1}^{\nu} b_j^{(n)} \sum_{i=1}^{\nu} a_{ji}^{(m)} \right] \mathcal{F}_m(t, y^0) \right) \Bigg] + O(\Delta t^3) \end{aligned} \tag{III.7}$$

1. Carl Gustav Jakob Jacobi (1804-1851) : mathématicien prussien

En comparant cette équation et le développement de Taylor (III.4), par identification, il existe $N(N+2)$ conditions pour assurer que le schéma soit globalement d'ordre 2 :

$$\mathbb{b}^{(n)} \cdot e = 1, \qquad \forall n \in [\![1; N]\!] \tag{III.8}$$

$$\mathbb{b}^{(n)} \cdot c^{(n)} = \frac{1}{2}, \qquad \forall n \in [\![1; N]\!] \tag{III.9}$$

$$\mathbb{b}^{(n)} \cdot \mathcal{A}^{(m)} \cdot e = \frac{1}{2}, \qquad \forall (n, m) \in [\![1; N]\!] \times [\![1, N]\!] \tag{III.10}$$

La condition (III.8) est celle portant sur la consistance du schéma, alors que les conditions (III.9) et (III.10) portent sur l'ordre 2 du schéma. Ainsi, il suffit que chaque schéma soit consistant pour que la méthode globale le soit. En revanche, pour que la méthode soit d'ordre 2, il n'est pas suffisant que chacun des schémas le soit. Notez que dans le cas où le système ne dépend pas explicitement du temps, nous pouvons définir :

$$c^{(n)} = \mathcal{A}^{(n)} \cdot e, \qquad \forall n \in [\![1; N]\!]$$

ce qui permet de respecter la condition (III.9), si la condition (III.10) est respectée.

Grâce à cette généralisation, nous sommes maintenant capable de développer de nombreux schémas RK-IMEX afin d'intégrer le système d'Euler. Mais avant d'imaginer de nouveaux schémas, il convient de reprendre l'étude de ceux déjà existants et dont les analyses n'ont pas été suffisamment profondes pour conclure sur leur viabilité dans un contexte opérationnel.

La contrainte HEVI pour les schémas IMEX à deux tableaux de Butcher

Historiquement, Ascher *et al.* (1995) [2, 3], proposaient d'utiliser les schémas IMEX pour résoudre des équations d'advections-diffusions pour lesquelles, les termes liés au processus de diffusions étaient traités de manière implicite et les termes advectifs étaient traités de manière explicite. Ce type de manipulation permet de s'affranchir de la contrainte CFL liée au nombre de Courant des processus rapides, générés par la diffusion. C'est donc une distinction sur la nature de processus qui a initialement motivée l'étude des méthodes IMEX. C'est exactement cette même démarche qui préside les fondements des schémas SI, où les processus rapides issus du système linéaire sont traités par un schéma implicite, dans toutes les directions.

Dans un contexte où la contrainte HEVI est imposée, la distinction ne se fonde plus seulement sur la nature des processus traités par tel ou tel schéma, mais également sur leurs directions de propagation. C'est pourquoi, les schémas IMEX HEVI s'affranchissent uniquement de la contrainte CFL des ondes rapides se propageant verticalement, tout en acceptant la contrainte sur la stabilité liée à la propagation horizontale des ondes acoustiques.

Des études préliminaires compilant de nombreux schémas RK-IMEX HEVI, utilisant deux tableaux de Butcher seulement et utilisés sur différents modèles, ont été réalisées (Pareschi & Russo (2005), Giraldo (2013) Ullrich & Jablonowski (2012) [25, 50, 67]). Certains de ces schémas sont d'ailleurs utilisés de manière opérationnelle, nonobstant le manque d'analyse de stabilité pour les systèmes complets d'Euler. Les études les plus complètes sur ces différents schémas RK-IMEX HEVI ont été pratiquées successivement par Weller *et al.* (2013) [68] et Lock *et al.* (2014) [42] qui introduisent deux versions différentes des schémas RK-IMEX :

- « UFpreF » (U-Forward et pression-Forward), où tous les termes liés à des opérateurs différentiels horizontaux sont calculés explicitement.
- « UFpreB » (U-Forward et pression-Backward), référence au schéma "forward-backward" de Mesinger (1977) où le gradient du vent est utilisé de manière implicite dans l'équation de la pression. Ce calcul se réalise par une simple substitution de la valeur du vent qui est calculée explicitement. Ainsi, aucune inversion horizontale n'est à déplorer.

Alors que pour les schémas d'Euler explicite/implicite du précédent chapitre, seuls les traitements avec une substitution implicite de l'horizontale permettaient d'avoir des schémas conditionnellement stables, Weller *et al.* (2013) [68] montrent que dans le cas des méthodes RK-IMEX HEVI la version UFpreF peut aussi être envisagée. Néanmoins, et Lock *et al.* (2014) [42] confirment cette tendance, il semble que, pour le système des ondes acoustiques 2D sans advection, la version avec un traitement implicite pour une partie horizontale soit plus stable que la version UFpreF, et ce, quel que soit le schéma. Il existe également une autre version implicite, déjà relevée dans le chapitre précédent, dont la littérature ne tient pas compte :

- « UBpreF » (U-Backward et pression-Forward), consistent à traiter le terme de pression avec une valeur explicite des termes horizontaux, et, une fois les inversions implicites réalisées, d'injecter la valeur des variables pour résoudre l'équation du mouvement horizontal implicitement.

Ce nouveau traitement a un avantage certain par rapport au UFpreB car il facilite grandement la résolution du système flot-dépendant, dont la coordonnée verticale varie dans le temps. Cet aspect sera mis en évidence dans le chapitre suivant. Il faut noter que ces trois versions restent envisageables avec une contrainte HEVI car, d'une part, elles n'engendrent aucune inversion sur la direction horizontale, et, de plus, elles ont numériquement le même coût. C'est pourquoi, les deux versions UFpreF et UFpreB sont systématiquement étudiées dans les travaux de Weller *et al.* (2013) [68] et Lock *et al.* (2014) [42], mais ceux-ci se cantonnent uniquement à utiliser ces schémas sur des systèmes très simplifiés. Comme nous l'avons constaté dans le précédent chapitre, les études de Skamarock & Klemp (1992) [60] ne sont pratiquées que sur le simple système des ondes acoustiques horizontales 1D, et l'étude sur le système complet révèle des zones d'instabilités supplémentaires dues à la direction verticale. Pour éviter cet écueil, Lock *et al.* (2014) [42] fondent leurs études sur le système d'ondes acoustiques 2D. Ceci permet de mettre en évidence certains comportements qui n'apparaissent pas sur le système encore plus simple du modèle de Skamarock & Klemp (1992) [60]. Mais ces études ne tiennent pas compte de l'advection qui semble déstabiliser fortement certains schémas, reconnus selon Lock *et al.* (2014) [42] comme étant les plus aptes à être utilisés dans un contexte opérationnel. C'est pour cette raison que nous proposons comme démarche de partir des études sur les systèmes les plus complets, et d'expliquer certains phénomènes à l'aide de modèles plus simples.

Les meilleurs schémas RK-IMEX à deux schémas, pour l'opérationnel, selon Lock *et al.* (2014) [42], sont ceux dont les tableaux de Butcher sont résumés ci-après :

UJ3(1,3,2) :

$$
\begin{array}{c|cccccc}
0 & 0 & & & & & \\
0 & 0 & 0 & & & & \\
1 & 0 & 1 & 0 & & & \\
1/2 & 0 & 1/4 & 1/4 & 0 & & \\
1 & 0 & 1/6 & 1/6 & 2/3 & 0 & \\
1 & 0 & 1/6 & 1/6 & 2/3 & 0 & 0 \\
\hline
 & 0 & 1/6 & 1/6 & 2/3 & 0 & 0
\end{array}
\qquad
\begin{array}{c|cccccc}
0 & 0 & & & & & \\
1/2 & 1/2 & 0 & & & & \\
1/2 & 1/2 & 0 & 0 & & & \\
1/2 & 1/2 & 0 & 0 & 0 & & \\
1/2 & 1/2 & 0 & 0 & 0 & 0 & \\
1 & 1/2 & 0 & 0 & 0 & 0 & 1/2 \\
\hline
1 & 1/2 & 0 & 0 & 0 & 0 & 1/2
\end{array}
$$

Ce schéma proposé par Ullrich & Jablonowski (2012) [67] jouit de plusieurs propriétés remarquables. C'est d'abord le seul proposant une unique inversion par pas de temps, ce qui entraîne une économie importante de calculs. De surcroît, il reste relativement peu onéreux, en dépit de la grandeur des tableaux. En effet, du fait que les trois dernières itérations explicites soient équivalentes, elles peuvent donc être mémorisées, ainsi que les quatre premières itérations implicites. Enfin, l'évaluation de (III.2) peut être évitée du fait que les lignes correspondantes au vecteur de poids sont égales aux lignes de la dernière évaluation (et donc $X^+ = X^{(\nu)}$). Par ailleurs, au-delà du coût raisonnable de ce schéma, il a été établi par Ullrich & Jablonowski (2012) [67] que la qualité des prévisions, offertes par ce schéma, était comparable à des schémas avec un ordre total plus élevé (ARS3(2,3,3) de Ascher *et al.* (1995) [3]). Nous pouvons remarquer que la partie explicite de ce schéma correspond au traditionnel Runge-Kutta-3 (ce qui explique l'ordre de précision de la partie explicite). Par ailleurs, ce schéma est déjà utilisé en opérationnel.

ARK2(2,3,2) :

$$
\begin{array}{c|ccc}
0 & 0 & & \\
2-\sqrt{2} & 2-\sqrt{2} & 0 & \\
1 & 1/2-\sqrt{2}/3 & 1/2+\sqrt{2}/3 & 0 \\
\hline
 & \sqrt{2}/2 & \sqrt{2}/2 & 1-\sqrt{2}/2
\end{array}
\qquad
\begin{array}{c|ccc}
0 & 0 & & \\
2-\sqrt{2} & 1-\sqrt{2}/2 & 1-\sqrt{2}/2 & \\
1 & \sqrt{2}/2 & \sqrt{2}/2 & 1-\sqrt{2}/2 \\
\hline
 & \sqrt{2}/2 & \sqrt{2}/2 & 1-\sqrt{2}/2
\end{array}
$$

Élaboré par Giraldo *et al.* (2013) [25], ce schéma a été montré comme étant le plus stable avec sa version UFpreB à la fois par Weller *et al.* (2013) [68] mais aussi par Lock *et al.* (2014) [42]. Comme

le précédent, ce schéma est aussi utilisé en opérationnel.

Trap2(2,3,2)(-1) :

0	0			
1	1	0		
1	1/2	1/2	0	
1	1/2	0	1/2	0
	1/2	0	1/2	0

0	0			
1	1	0		
1	1/2	0	1/2	
1	1/2	0	0	1/2
	1/2	0	0	1/2

De manière générale, les schémas Runge-Kutta ayant un vecteur d'avance temporelle toujours égal à 1, et une combinaison finale correspondant au calcul de la dernière étape, peuvent être considérés comme des schémas de prédiction-correction (et réciproquement). Ce schéma est donc un simple schéma de prédiction-correction avec une première étape purement explicite et les deux suivantes utilisant un schéma trapézoïdal (itérations explicites pour l'un et implicites pour l'autre). Lock *et al.* (2014) [42] étudient ce schéma ainsi qu'une variante nommée Trap2(2,3,2) qui diffère simplement de celui-ci en enlevant l'évaluation explicite de la partie implicite lors du schéma prédictif. Bien que leurs études concluent à des propriétés similaires entre ces deux schémas, nous proposons ici de ne garder que Trap2(2,3,2)(-1) qui est bien plus stable en présence d'advections.

Les schémas SSP (établis par Pareschi & Russo (2005) [50]) sont directement écartés car ils sont à la fois plus chers et moins stables que les autres candidats retenus. De même, les schémas ARS de Ascher *et al.* (1997) [2] sont aussi précis, mais moins stables, et pour certains, également plus coûteux. C'est pourquoi, la liste de nos candidats se réduit seulement aux trois schémas ci-avants.

Maintenant que la liste des premiers candidats est définie, il est possible de présenter les résultats des études de stabilité sur le système linéaire, ainsi que la mise en œuvre des algorithmes de résolution. Pour les systèmes non-linéaires, le sujet sera traité au chapitre suivant.

2 Étude de stabilité présence d'advection

Dans le cas des études des schémas RK-IMEX HEVI utilisant uniquement deux schémas, nous suivons la démarche présentée dans le chapitre précédent, nous découpons l'opérateur linéaire sans bord $l\mathscr{L}$ en deux parties. La partie $\mathscr{I}$ est composée des termes d'ajustement verticaux (ceux responsables de la propagation verticale des ondes acoustiques), c'est-à-dire la dérivée verticale de la pression dans l'équation du mouvement vertical (lL2), ainsi que la dérivée verticale du vent vertical dans les équations de la température (lL3) et de la pression (lL4). Soit également la partie $\mathscr{E}$ des termes horizontaux traités de manière explicite (comme l'advection). Le reste des termes du système, liés à l'ajustement horizontal, est réparti entre les deux parties avec les clés δ_u et δ_p afin de pouvoir réaliser les trois versions de ces schémas : UFpreF, UFpreB et UBpreF. Pour avoir

UFpreF	UFpreB	UBpreF
$\delta_u = \delta_p = 0$	$\delta_u = 0\ ;\ \delta_p = 1$	$\delta_u = 1\ ;\ \delta_p = 0$

TABLE III.1 – *Valeurs des coefficients δ_u et δ_p pour produire les trois versions différentes de traitement HEVI.*

un schéma explicite sur l'horizontale, il suffit de ne pas avoir d'inversion implicite à effectuer pour la direction horizontale. Ceci n'est pas contradictoire avec le fait que certains termes horizontaux soient traités de manière implicite, comme il en a déjà été question. En reprenant les notations des précédents chapitres, les parties $\mathscr{I}$ et $\mathscr{E}$ se résument à deux matrices 4×4 :

$$\mathbf{I} = -\begin{pmatrix} 0 & 0 & -\delta_u R\hat{\imath}k & \delta_u R\overline{T}\hat{\imath}k(\hat{\imath}\ell(r) + \frac{1}{2}) \\ 0 & 0 & 0 & g(\hat{\imath}\ell(r) + \frac{1}{2}) \\ \delta_p \frac{R\overline{T}}{C_v}\hat{\imath}k & \frac{R\overline{T}}{C_v\overline{H}}(\hat{\imath}\ell(r) - \frac{1}{2}) & 0 & 0 \\ \delta_p \hat{\imath}k(\frac{C_p}{C_v}((\hat{\imath}\ell(r) + \frac{1}{2}) - 1) & -\frac{C_p}{C_v}\overline{H}(\ell(r)^2 + \frac{1}{4}) & 0 & 0 \end{pmatrix}$$

$$\mathbf{E} = -\begin{pmatrix} \overline{U}\hat{\imath}k & 0 & -(1-\delta_u)R\hat{\imath}k & (1-\delta_u)R\overline{T}\hat{\imath}k(\hat{\imath}\ell(r) + \frac{1}{2}) \\ 0 & \overline{U}\hat{\imath}k & 0 & 0 \\ (1-\delta_p)\frac{R\overline{T}}{C_v}\hat{\imath}k & 0 & \overline{U}\hat{\imath}k & 0 \\ (1-\delta_p)\hat{\imath}k(\frac{C_p}{C_v}((\hat{\imath}\ell(r) + \frac{1}{2}) - 1) & 0 & 0 & \overline{U}\hat{\imath}k \end{pmatrix}$$

avec δ_u et δ_p valant soit 0 soit 1, sachant que la contrainte HEVI, impose que δ_u et δ_p ne peuvent simultanément valoir 1. Le Tableau III.1 résume la valeur des coefficientsδ_u et δ_p pour obtenir ces trois versions différentes de traitement HEVI.

Il est maintenant possible de réaliser l'analyse de stabilité issue des techniques décrites par Von Neumann. Quels que soient le schéma ou la version utilisés, la matrice d'amplification se définit à partir de la discrétisation (III.2)-(III.3) :

$$A^{(j)} = \mathrm{I} + \Delta t \sum_{i=1}^{j-1} \tilde{a}_{ji}\mathbf{E} \cdot A^{(i)} + \sum_{i=1}^{j} a_{ji}\mathbf{I} \cdot A^{(i)} \tag{III.11}$$

$$A = \mathrm{I} + \Delta t \sum_{j=1}^{\nu} \tilde{b}_{j}\mathbf{E} \cdot A^{(j)} + \sum_{j=1}^{\nu} b_{j}\mathbf{I} \cdot A^{(j)} \tag{III.12}$$

A est donc une matrice complexe 4×4, qui possède quatre valeurs propres complexes λ_l avec $l \in [\![1; 4]\!]$. L'amplification et la phase de l'onde considérée sont données respectivement par le module $\Gamma_l = |\lambda_l|$ et l'argument $\theta_l = \mathrm{Arg}(\lambda_l)$ de la valeur propre λ_l. En rappelant que la vitesse d'advection est calculée par rapport au nombre de Mach M_U, alors chaque valeur propre dépend à la fois de C_*, r et M_U, et la stabilité globale du schéma est définie par rapport au coefficient d'amplification Γ :

$$\Gamma(C_*, r, M_U) = \max_{l \in [\![1;4]\!]} \{\Gamma_l(C_*, r, M_U)\}$$

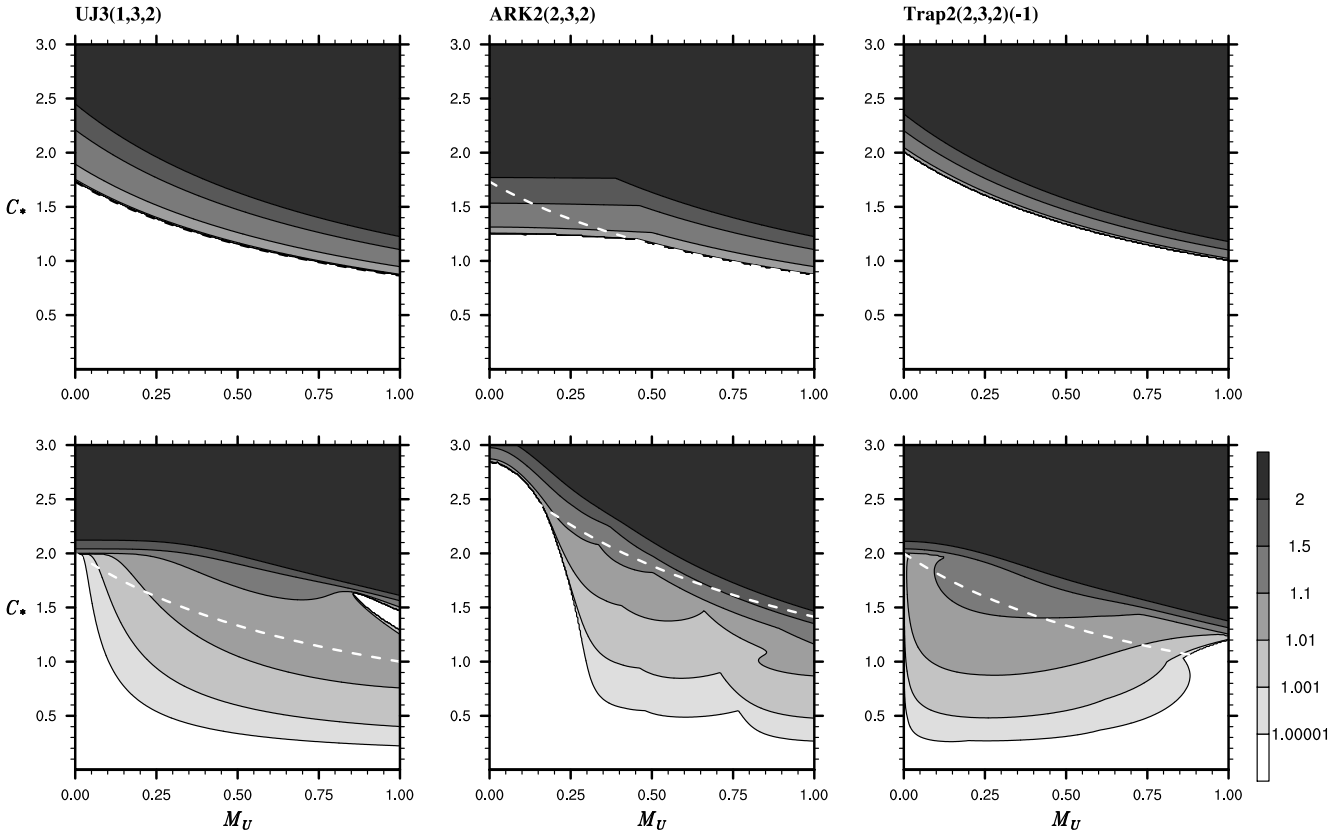


FIGURE III.1 – *Coefficient d'amplification* Γ_r *en fonction de* C_* *et* M_U *pour les trois schémas en version UFpreF (en haut) et UFpreB (en bas).*

Ainsi, le schéma est stable si et seulement si $\Gamma \leq 1$.

Il faut cependant noter une chose : pour les études de stabilité linéaire, les versions UFpreB et UBpreF sont rigoureusement identiques. Par souci de lisibilité, nous proposons donc de ne présenter que les résultats opposant UFpreF à UFpreB et de fixer donc $\delta_u = 0$.

La présence d'advection modifie de manière importante les phases des solutions numériques. Dans ce cas, il est difficile de faire la distinction entre les ondes et faire leur étude séparée. C'est pourquoi dans cette partie, nous n'étudions que la stabilité globale en posant :

$$\Gamma_r(C_*, M_U) = \max_{r \in [10^{-2}; 10^3]} \{\Gamma(C_*, r, M_U)\}$$

Là encore, la stabilité du schéma est équivalente à $\Gamma_r \leq 1$.

La Figure III.1 montre le coefficient d'amplification maximale pour r variant dans l'intervalle $[10^{-2}, 10^3]$ de l'ensemble des schémas en fonction de C_* et M_U dans les deux versions UFpreF et UFpreB. Les zones blanches correspondent à une valeur du coefficient d'amplification inférieure ou égale à 1. Ceci signifie qu'avec ces paramètres C_* et M_U, le schéma est stable, quelle que soit la valeur de r variant de $[10^{-2}, 10^3]$. À l'inverse, les zones grisées correspondent à des jeux de variables C_* et M_U tels que il existe au moins un mode défini par r tel que les schémas sont instables. Nous observons que le nombre de Courant critique des schémas UJ3(1,3,2) et Trap2(2,3,2)(-1) semble

décrire dans les deux cas une branche hyperbolique. Pour le cas ARK2(2,3,2), dès que le nombre de Mach est supérieur à 1/2, alors la stabilité de ce schéma semble identique à UJ3(1,3,2). Ce comportement est dû au fait que, au repos, la stabilité de ARK2(2,3,2) décroît en fonction de r car, pour de petites valeurs de ce rapport, le nombre de Courant critique est $\sqrt{3}$. En revanche, dès que r devient grand (supérieur à 100), la stabilité diminue pour arriver à la valeur de 1,25 (*cf* Lock *et al.* (2014) [42]). Dans le cas du ARK2(2,3,2), il semble que la contrainte CFL soit composée de deux conditions : d'une part, il est nécessaire que $C_* \leq 1{,}25$ (comme l'impose les études sur le système au repos) et d'autre part, la même contrainte que UJ3(1,3,2). Pour vérifier cette hypothèse, nous avons baissé la valeur maximale de r en la fixant à 10^{-1}, et dans ce cas, nous avons exactement les mêmes courbes que pour le schéma de UJ3(1,3,2).

L'étude sur le système simplifié des ondes acoustiques 1D horizontales permet d'établir simplement les relations hyperboliques décrites par le nombre de Courant critique. En effet, ce dernier doit vérifier :

$$(1 + |M_U|)C_* \leq \frac{\sqrt{2(\tilde{\mathfrak{b}} \cdot \tilde{\mathcal{A}} \cdot \tilde{c}) - 1/4}}{\tilde{\mathfrak{b}} \cdot \tilde{\mathcal{A}} \cdot \tilde{c}} \tag{III.13}$$

avec, pour le cas UJ3(1,3,2) et ARK2(2,3,2), $\tilde{\mathfrak{b}} \cdot \tilde{\mathcal{A}} \cdot \tilde{c} = 1/6$ (et donc $(1 + |M_U|)C_* \leq \sqrt{3}$), et pour Trap2(2,3,2)(-1) $\tilde{\mathfrak{b}} \cdot \tilde{\mathcal{A}} \cdot \tilde{c} = 1/4$ (soit $(1 + |M_U|)C_* \leq \sqrt{2}$).

La leçon la plus importante à retenir de cette étude est le comportement très instable de ces schémas en version UFpreB dès lors qu'il existe un écoulement. La ligne pointillée blanche représente la branche hyperbolique telle que pour M_U nul, la valeur du graphe soit celle correspondant au nombre de Courant maximum. Alors que l'ensemble de la littérature laissait croire, grâce aux études menées sur les systèmes au repos, que les versions UFpreB étaient plus stables que les versions explicites UFpreF, il apparaît clairement que la stabilité est plus faible que la valeur attendue par la relation hyperbolique. Pire encore, les schémas UJ3(1,3,2) et Trap2(2,3,2)(-1) semblent inconditionnellement instables en présence d'advection. Seul le schéma ARK2(2,3,2) reste stable, mais uniquement pour des valeurs du nombre de Mach inférieures à 0,3.

Les instabilités apparaissant dans le cas des versions UFpreB (mais aussi UBpreF) se trouvent également pour le simple système linéaire acoustique 1D horizontal. La matrice d'amplification de la version UFpreB s'écrit sous la somme de la matrice d'amplification de la version UFpreF avec en plus des termes résiduels d'ordre encore supérieur en Δt. Ce sont ces résidus qui sont responsables de la modification importante de la stabilité des schémas. Nous avons observé que la dégradation de la stabilité était liée à quelques traitements implicites. Une approche heuristique d'annulation progressive des termes de la matrice d'amplification a permis de mettre en évidence la responsabilité des termes en Δt^5. Ces manipulations *ad-hoc* des tableaux de Butcher ont permis d'émettre une hypothèse sur la source de ces instabilités. Il est clair que la version UFpreF sauvegarde la structure des caractéristiques car elles sont toutes traitées par le même schéma. En effet, la version UFpreB impose un traitement implicite de la divergence horizontale, et brise ces caractéristiques. De même pour la version UBpreF. Pour pallier cette difficulté, il semble nécessaire de rétablir cet équilibre par un traitement implicite successif des termes d'ajustements horizontaux. Cette étude semble pointer les faiblesses d'une stratégie RK-IMEX HEVI uniquement à deux tableaux de Butcher. C'est pourquoi, et grâce à la démonstration précédente, il est possible d'envisager l'introduction

de schémas supplémentaires afin de traiter séparément les termes d'advections, d'ajustements verticaux et les termes d'ajustements horizontaux. Cette analyse plaide en faveur de l'introduction de plusieurs tableaux de Butcher pour traiter chacun des termes affectant la stabilité globale du schéma.

3 Proposition de schéma RK-IMEX HEVI à quatre tableaux de Butcher

Une nouvelle alternative de schéma RK-IMX HEVI utilisant quatre tableaux de Butcher est proposée ci-dessous. Dans cette nouvelle voie, nous exploitons le fait que les termes d'ajustement ($\nabla \cdot \mathbf{V}$ et ∇p) peuvent être alternativement traités soit explicitement, soit implicitement. En effet, tant que ces deux termes ne sont pas traités simultanément de manière implicite, alors, aucune inversion horizontale n'est à effectuer, et donc, nous respectons bien la contrainte HEVI. Le principe est d'isoler ces termes et de leur associer un traitement spécifique. Symboliquement, nous pouvons écrire :

$$\partial_t X = \mathscr{E}'(X) + \mathscr{I}'(X) + \mathscr{U}(X) + \mathscr{P}(X) \tag{III.14}$$

où $\mathscr{E}'$ contient l'ensemble des termes advectifs et, de manière générale, l'ensemble des termes devant être traité de manière explicite. $\mathscr{I}'$ contient uniquement les termes d'ajustements verticaux, $\mathscr{U}$ contient l'ensemble des termes de divergences horizontales, et $\mathscr{P}$ le gradient de pression horizontal.

Énoncé ainsi, il apparaît clairement que le système peut s'intégrer via l'utilisation de quatre tableaux de Butcher chacun associé à une des parties définies ci-avant. $\mathscr{E}'$ et $\mathscr{I}'$ sont intégrés par les tableaux précédemment définis $\{\tilde{c}, \tilde{\mathcal{A}}, \tilde{b}\}$ et $\{c, \mathcal{A}, b\}$, alors que les contributions $\mathscr{U}$ et $\mathscr{P}$ sont respectivement intégrées par les tableaux de Butcher $\{c^u, \mathcal{A}^u, b^u\}$ et $\{c^p, \mathcal{A}^p, b^p\}$, qui sont non-nécessairement explicites. Afin de respecter la contrainte HEVI, les coefficients diagonaux de ces deux derniers tableaux ne peuvent simultanément être non-nuls :

$$a^u_{jj} a^p_{jj} = 0 \quad \text{pour} \quad j \in [\![1; \nu]\!] \tag{III.15}$$

La construction des schémas à quatre tableaux de Butcher peut se réaliser en prenant un schéma RK-IMEX original pour lequel nous alternons à chaque étape entre la version UFpreB et UBpreF. Autrement dit, considérons la j-ième et la $(j+1)$-ième étape (j allant de 1 à $\nu-1$). Le terme inclus dans $\mathscr{P}$ est calculé selon la j-ième étape du tableau explicite (ou implicite) alors que le terme inclus dans $\mathscr{U}$ est calculé selon la j-ième étape du tableau implicite (respectivement explicite), alors que pour l'étape suivante, le terme dans $\mathscr{U}$ est traité par le tableau explicite (respectivement implicite) et le terme de $\mathscr{P}$ est calculé par le tableau implicite (respectivement explicite). Par exemple, pour

le schéma original du ARK2(2,3,2), cela conduit aux quatre tableaux de Butcher suivants :

$\{\tilde{c}, \tilde{\mathcal{A}}, \tilde{b}\}$

0	0		
$2-\sqrt{2}$	$2-\sqrt{2}$	0	
1	$1/2-\sqrt{2}/3$	$1/2+\sqrt{2}/3$	0
	$\sqrt{2}/2$	$\sqrt{2}/2$	$1-\sqrt{2}/2$

$\{c, \mathcal{A}, b\}$

0	0		
$2-\sqrt{2}$	$1-\sqrt{2}/2$	$1-\sqrt{2}/2$	
1	$\sqrt{2}/2$	$\sqrt{2}/2$	$1-\sqrt{2}/2$
	$\sqrt{2}/2$	$\sqrt{2}/2$	$1-\sqrt{2}/2$

$\{c^u, \mathcal{A}^u, b^u\}$

0	0		
$2-\sqrt{2}$	$2-\sqrt{2}$	0	
1	$\sqrt{2}/2$	$\sqrt{2}/2$	$1-\sqrt{2}/2$
	$\sqrt{2}/2$	$\sqrt{2}/2$	$1-\sqrt{2}/2$

$\{c^p, \mathcal{A}^p, b^p\}$

0	0		
$2-\sqrt{2}$	$1-\sqrt{2}/2$	$1-\sqrt{2}/2$	
1	$1/2-\sqrt{2}/3$	$1/2+\sqrt{2}/3$	0
	$\sqrt{2}/2$	$\sqrt{2}/2$	$1-\sqrt{2}/2$

Et pour le Trap2(2,3,2)(-1) :

$\{\tilde{c}, \tilde{\mathcal{A}}, \tilde{b}\}$

0	0			
1	1	0		
1	$1/2$	$1/2$	0	
1	$1/2$	0	$1/2$	0
	$1/2$	0	$1/2$	0

$\{c, \mathcal{A}, b\}$

0	0			
1	1	0		
1	$1/2$	0	$1/2$	
1	$1/2$	0	0	$1/2$
	$1/2$	0	0	$1/2$

$\{c^u, \mathcal{A}^u, b^u\}$

0	0			
1	1	0		
1	$1/2$	0	$1/2$	
1	$1/2$	0	$1/2$	0
	$1/2$	0	$1/2$	0

$\{c^p, \mathcal{A}^p, b^p\}$

0	0			
1	1	0		
1	$1/2$	$1/2$	0	
1	$1/2$	0	0	$1/2$
	$1/2$	0	0	$1/2$

Il est facile de montrer que ces tableaux vérifient la condition d'ordre 2 du schéma global (III.8)-(III.10), démontrée précédemment. De plus, ces nouveaux schémas possèdent les propriétés principales de leurs originaux à savoir qu'il est d'ordre explicite 2 (car il y a le même tableau de Butcher pour la partie explicite que pour l'original), le même nombre d'inversions, le même coefficient de stockage, et enfin le même ordre global. Ainsi, nous appelons ces nouveaux schémas respectivement ARK2-Mixed et Trap2-Mixed. Néanmoins, il n'est pas certain que le schéma ARK2-Mixed conserve la L-stabilité.

Ce principe de construction a été mis à l'épreuve dans le cas du schéma UJ3(1,3,2), mais ceci s'est révélé décevant car ce nouveau schéma, bien que stable, n'est que d'ordre 1 pour la précision globale et devient inenvisageable pour la PNT. Pour pallier cette difficulté, une recherche plus approfondie a permis d'élaborer un schéma d'ordre 2, à partir des tableaux initiaux de UJ3(1,3,2). Ce nouveau schéma, appelé UJ3-Mixed, possède également les propriétés essentielles du schéma original, notamment le fait qu'il ne possède qu'une itération implicite. Il se présente ainsi :

$\{\tilde{c}, \tilde{\mathcal{A}}, \tilde{b}\}$

0	0					
0	0	0				
1	0	1	0			
1/2	0	1/4	1/4	0		
1	0	1/6	1/6	2/3	0	
1	0	1/6	1/6	2/3	0	0
	0	1/6	1/6	2/3	0	0

$\{c, \mathcal{A}, b\}$

0	0					
1/2	1/2	0				
1/2	1/2	0	0			
1/2	1/2	0	0	0		
1/2	1/2	0	0	0	0	
1	1/2	0	0	0	0	1/2
1	1/2	0	0	0	0	1/2

$\{c^u, \mathcal{A}^u, b^u\}$

0	0					
1/2	1/2	0				
1/2	1/2	0	0			
1/2	0	1/4	1/4	0		
1	0	1/6	1/6	2/3	0	
1	0	1/6	1/6	2/3	0	
1	0	1/6	1/6	2/3	0	

$\{c^p, \mathcal{A}^p, b^p\}$

0	0					
0	0	0				
1	0	1	0			
1/2	0	1/4	1/4	0		
1	0	1/6	1/6	2/3	0	
1	1/2	0	0	0	0	1/2
1	1/2	0	0	0	0	1/2

L'étude de la stabilité de ce schéma en présence d'advection est réalisée de manière analogue à celle des autres schémas, et est illustrée par la Figure III.2. Nous pouvons remarquer que, pour le cas de ARK2-Mixed, l'introduction de deux autres tableaux de Butcher améliore que faiblement la stabilité du schéma par rapport à la version UFpreF. En effet, mis à part pour une vitesse d'advection inférieure à un nombre de Mach proche de 0,25, nous constatons que la stabilité est vraiment similaire à la version UFpreF. Ainsi, ce changement ne corrige pas toutes les instabilités de la version UFpreB car il ne permet pas d'avoir un nombre de Courant au-delà de $\sqrt{3}$. En revanche, ce traitement semble corriger l'inconditionnelle instabilité du schéma ARK2-Mixed original pour un vent supérieur à 0,5 Mach. Malgré nos efforts, il semble donc que ce nouveau schéma reste comparable à la version originelle explicite UFpreF.

Dans le cas de UJ3-Mixed et Trap2-Mixed, les conclusions sont remarquables. D'une part, ce traitement corrige totalement l'inconditionnelle instabilité des versions UFpreB en présence d'advection

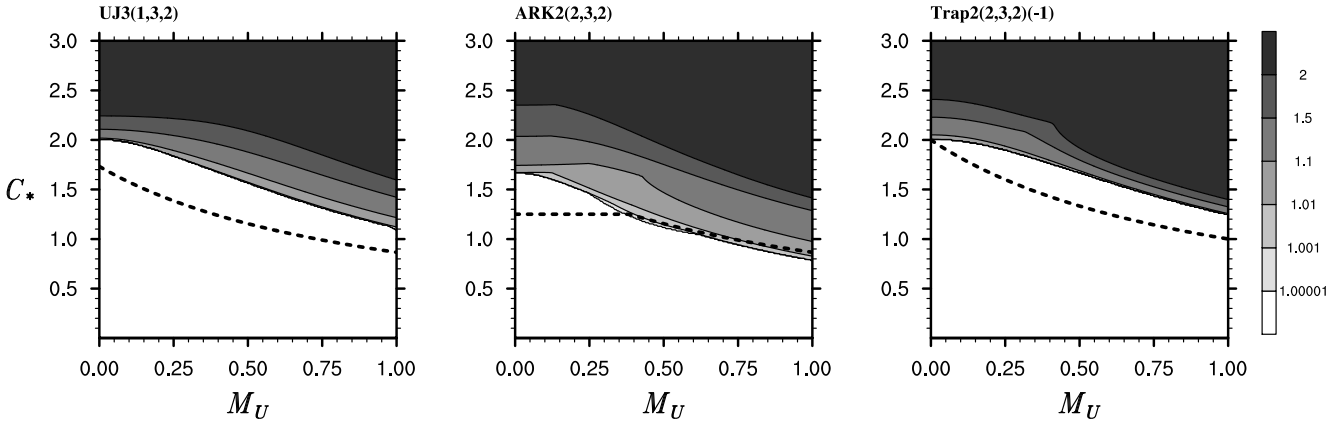


FIGURE III.2 – *Même graphique que Figure III.1, mais pour les schémas UJ3-Mixed (à gauche), ARK2-Mixed (au milieu) et Trap2-Mixed (à droite). Les lignes en pointillés rappellent le nombre maximal de Courant des versions UFpreF.*

et ont la même stabilité que cette version au repos ($M_U = 0$), mais en plus, ces nouveaux schémas sont plus stables que leurs versions UFpreF (rappelées par la ligne en pointillées noire).

La nouvelle classe de schémas que nous avons définie nous a permis d'introduire de nouveaux schémas qui sont plus stables que ceux présents dans la littérature et à coût constant. En effet, comme les deux schémas que nous introduisons ont la même taille que les tableaux initiaux, aucune mémoire supplémentaire n'est nécessaire. Ce gain de stabilité revient à respecter un ordre de calculs particulier, sans en nécessiter d'avantage.

À ce stade des études, il semble que les meilleurs candidats soient le UJ3-Mixed et Trap2$(2,3,2)(-1)$ dans sa version UFpreF (à deux tableaux de Butcher) ainsi que le Trap2-Mixed. Reste à examiner leurs impacts sur les modes acoustiques et de gravité.

4 Étude des erreurs de phases acoustiques et de gravité

L'approximation des solutions réalisées par une intégration numérique des solutions peut déformer les solutions ondulatoires de deux manières. Soit elle modifie l'amplitude du signal (amortissement ou amplification dans le cas instable) soit la vitesse (en ralentissant ou en accélérant). Alors que de faibles différences des phases peuvent ne pas nuire à la qualité du schéma, de fortes variations peuvent, quant à elles, poser de plus lourds problèmes. En effet, si la vitesse de groupe (dérivée de la fréquence par rapport au nombre d'ondes) s'annule, cela signifie que la propagation d'un paquet d'ondes ne peut être représentée par le schéma, et donc, qu'il peut y avoir localement des accumulations d'énergie qui, au mieux, génèrent du bruit, et au pire, rendent le schéma instable.

Afin d'assurer la non-déformation des ondes de gravité, qui doivent être représentées de manière la plus exacte possible, nous allons réaliser la même étude que Lock *et al.* (2014) en prenant le soin de bien distinguer, parmi les valeurs propres λ_l (avec $l \in [\![1;4]\!]$), lesquelles s'associent aux ondes acoustiques et celles des ondes de gravité. Après quoi, nous pourrons observer séparément les deux

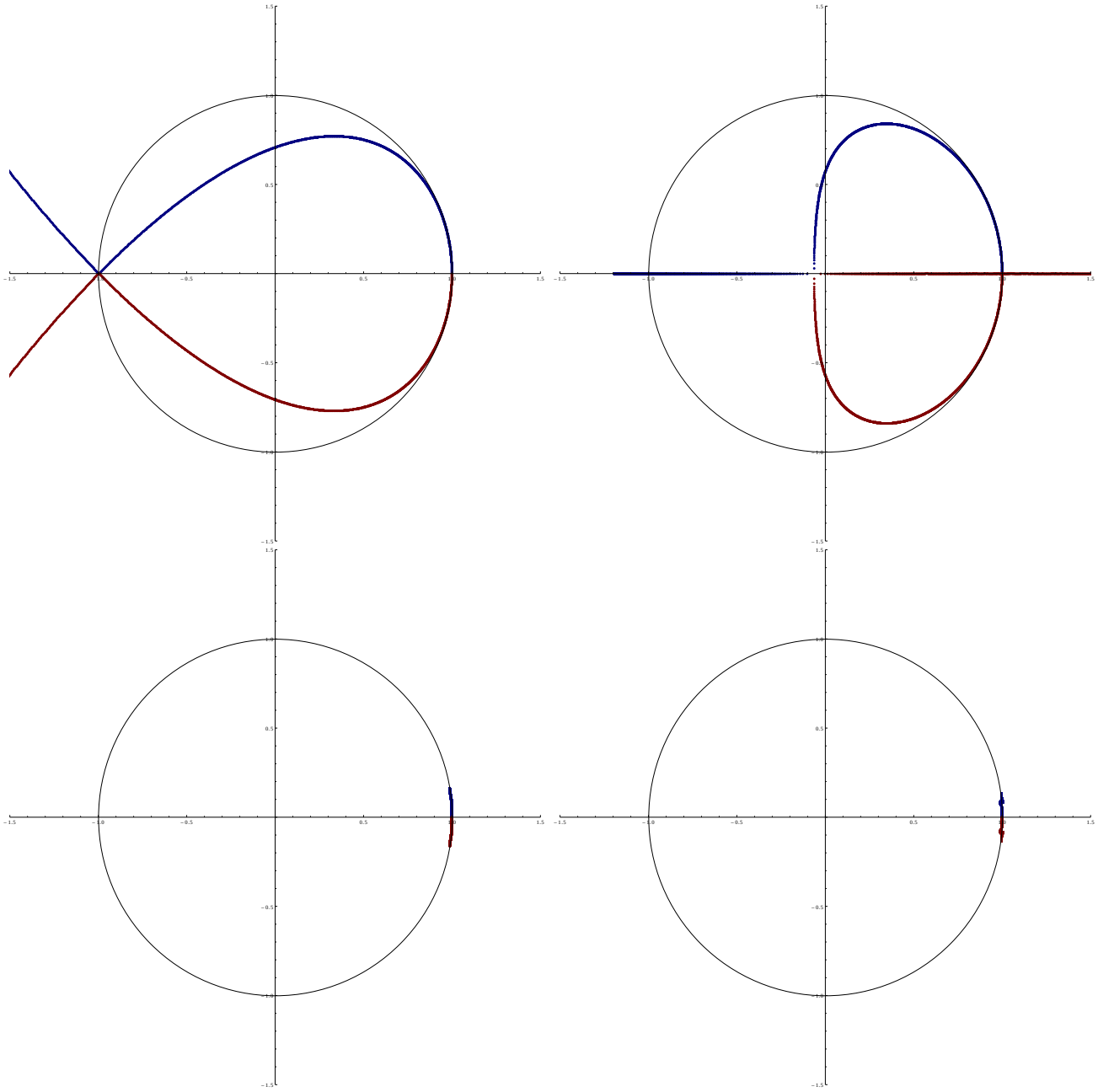


FIGURE III.3 – *Valeurs propres complexes de la matrice d'amplification du schéma Trap2(2,3,2)(-1) sous la version UFpreF (à gauche)* $r = 0{,}1$ *et Mixed (à droite) pour* $r = 0{,}7$ *et pour plusieurs valeurs de* C_**. En haut, nous représentons les deux ondes acoustiques et en bas les deux ondes de gravité.*

processus, afin de voir leurs amortissements respectifs, ainsi que leurs phases. Avant tout, il s'agit d'évoquer la manière dont nous avons procédé en ce qui concerne la distinction des deux ondes. Il faut noter que pour avoir la possibilité de pratiquer ces associations par paire, nous allons procéder par l'étude des phases de chaque onde. Or, pour faire émerger certaines propriétés sur ces phases,

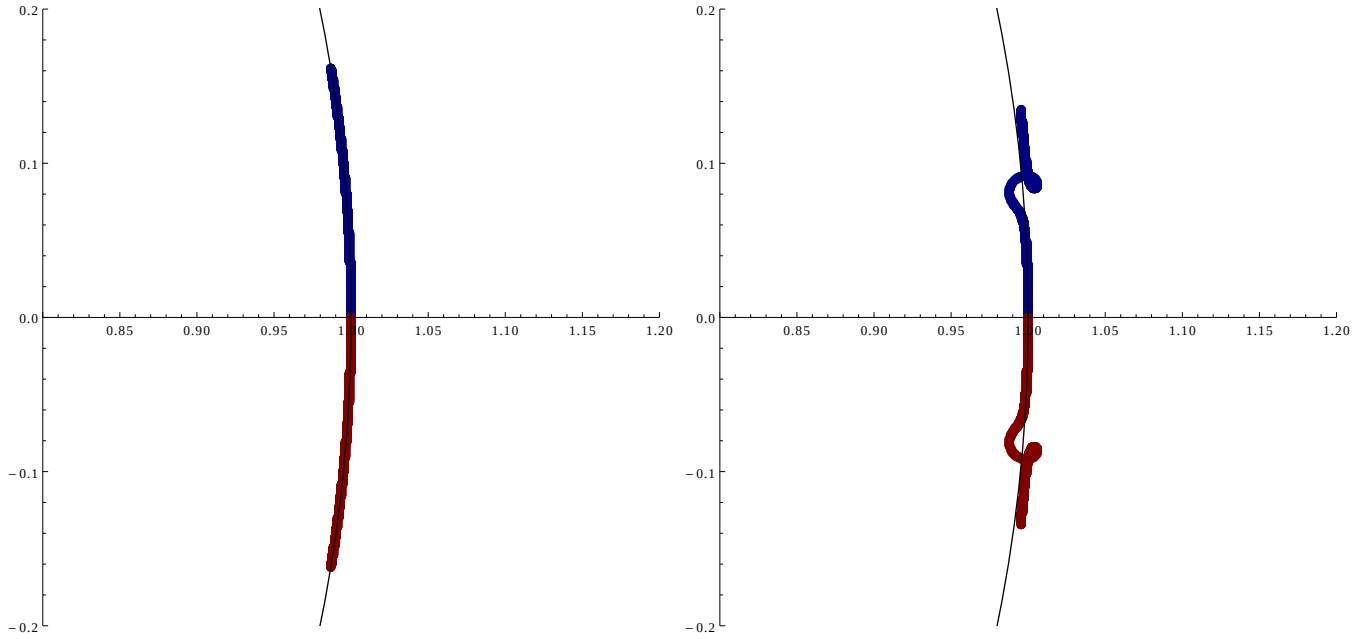


FIGURE III.4 – *Même graphique que Figure III.3 avec un agrandissement autour de 1 pour illustrer les instabilités des ondes de gravité.*

il est nécessaire que le système soit au repos. Dans le reste du chapitre, nous imposons donc $\overline{U} = 0$.

Algorithme de discrimination des ondes

Nous savons que les valeurs propres λ_l définies par les schémas d'intégration temporels doivent être proches (du moins pour un faible nombre de Courant) des solutions du système qui sont des ondes. C'est pourquoi, les valeurs propres des matrices d'amplification sont, dans la plupart des cas, deux paires de valeurs complexes conjuguées. Cette propriété montre donc bien que les ondes numériques se propagent à la même vitesse, dans la même direction, mais dans un sens opposé. Ainsi, il est facile d'isoler les deux paires. Reste à savoir quelle est celle se rapprochant des ondes acoustiques, et celle des ondes de gravité. Pour cela, il suffit de faire une évaluation des phases : la paire ayant la plus grande (en valeur absolue) est celle faisant référence aux ondes acoustiques, et l'autre celle qui renvoie aux ondes de gravité.

Il faut noter que dans le cas des ondes acoustiques, cette opposition entre les phases n'est pas toujours respectée. En effet, la Figure III.3 montre, sur le plan complexe, la position de plusieurs valeurs propres pour un ℓ fixé, en fonction de C_* allant de $[10^{-3}, 3]$ après avoir séparé les ondes de gravité et les ondes acoustiques. Nous voyons qu'il existe des cas où les valeurs propres, associées à une paire, deviennent réelles, avec la valeur de leurs phases valant π et 0. Dans ce cas précis, nous faisons jouer l'argument de continuité sur les valeurs propres en fonction de C_*. De là, nous observons que la paire qui reste complexe prolonge dans la continuité la valeur des ondes de gravité. Nous concluons donc que la paire réelle correspond aux ondes acoustiques. Dans ce cas-là, et le schéma Trap-Mixed illustre parfaitement ce comportement (Figure III.3), il est très difficile de

différencier l'onde positive de l'onde négative. Dans ce cas, un choix arbitraire de notre part sera effectué. Cet algorithme est résumé dans l'Annexe B.

Malgré certains cas pathologiques pour les ondes acoustiques, les exemples illustrés par la Figure III.3 prouve que cet algorithme associe correctement les valeurs propres entre elles afin d'assurer la continuité de leurs comportements en fonction des variables C_* et r. Le schéma est stable tant que l'ensemble des valeurs propres se trouvent dans le disque unité. En effet, dans le cas contraire, cela signifie qu'il existe au moins une valeur propre pour laquelle le module est supérieur à 1, et donc Γ est plus grand que 1. Forts de cet outil, nous pouvons maintenant décrire le comportement respectif de chaque onde.

Comportement spécifique de chaque ondes sans advection

La Figure III.5 décrit séparément le comportement critique à la fois des ondes acoustiques et des ondes de gravité pour le schéma UJ3(3,1,2) dans sa version UFpreF. Pour chaque onde, la figure de gauche montre le coefficient d'amplification maximal si l'onde est instable, et le minimal dans le cas stable pour mettre en exergue l'amortissement créé par ces schémas. La figure de droite montre la phase numérique positive de l'onde considérée (donc variant de $[0, \pi]$) uniquement dans le cas où la stabilité globale du schéma est respectée. Dans le cas contraire, ces zones sont blanches. Les zones noires des graphiques sur les phases signifient que la valeur de la phase est π, et donc, les valeurs propres des ondes associées sont réelles. Enfin, comme les ondes de gravité ont une phase bien plus faible que les ondes acoustiques (*cf* Figure III.3), nous avons multiplié l'argument de la valeur propre par 32.

Afin d'être capable d'interpréter ces figures, il convient d'établir une remarque préliminaire. Il faut bien comprendre qu'un modèle capable de modéliser une onde décrite par r l'est également pour un ratio r plus petit. C'est pourquoi, le critère de stabilité CFL se déduit comme étant la plus petite valeur atteinte par C_* pour toutes les valeurs de r et quelles que soient les ondes décrites par ce schéma. Ainsi, dans le cas général, la stabilité peut décroître en fonction de r, et ce, malgré le traitement implicite de la verticale. Il est important de garder à l'esprit que les résultats que nous montrons ici ne sont valables que pour les valeurs de r que nous utilisons.

La Figure III.5 montre que les instabilités du schéma appliqué à ce système proviennent des deux ondes. Nous retrouvons la limite $C_* \leq \sqrt{3}$ imposée par le schéma Runge-Kutta-3. De plus, il existe une zone comprise entre $r > 1$ et $C_* > 0{,}75$ telle que les ondes de gravité sont sensiblement amorties. Le déformations qui semblent les plus importantes sont dues aux erreurs de phase des deux ondes. En effet, pour le cas des ondes acoustiques, il existe une zone entre $r > 100$ et $C_* > 0{,}75$ telle que la phase est constante (égale à π). Or, si la phase est constante, cela implique que la vitesse de groupe s'annule pour de nombreuses ondes, avec les risques que nous avons déjà indiqués. Il semble également exister certaines valeurs de r autour de 2 telle qu'il existe une inversion de phase pour les ondes de gravité, avec également un risque d'annulation de la vitesse de groupe.

La Figure III.6 confirme clairement que la stabilité de ce schéma Trap2(2,3,2(-1) UFpreF est conditionné par les ondes acoustiques, et impose, pour ce système au repos, $C_* \leq 2$. De plus, il possède une très bonne représentation des ondes de gravité, en maintenant neutre l'amplitude du signal, et sans ralentissement ni accélération. En revanche pour les ondes acoustiques, il existe une plage relativement importante où l'une des ondes acoustiques est fortement amortie ($r > 10$ et

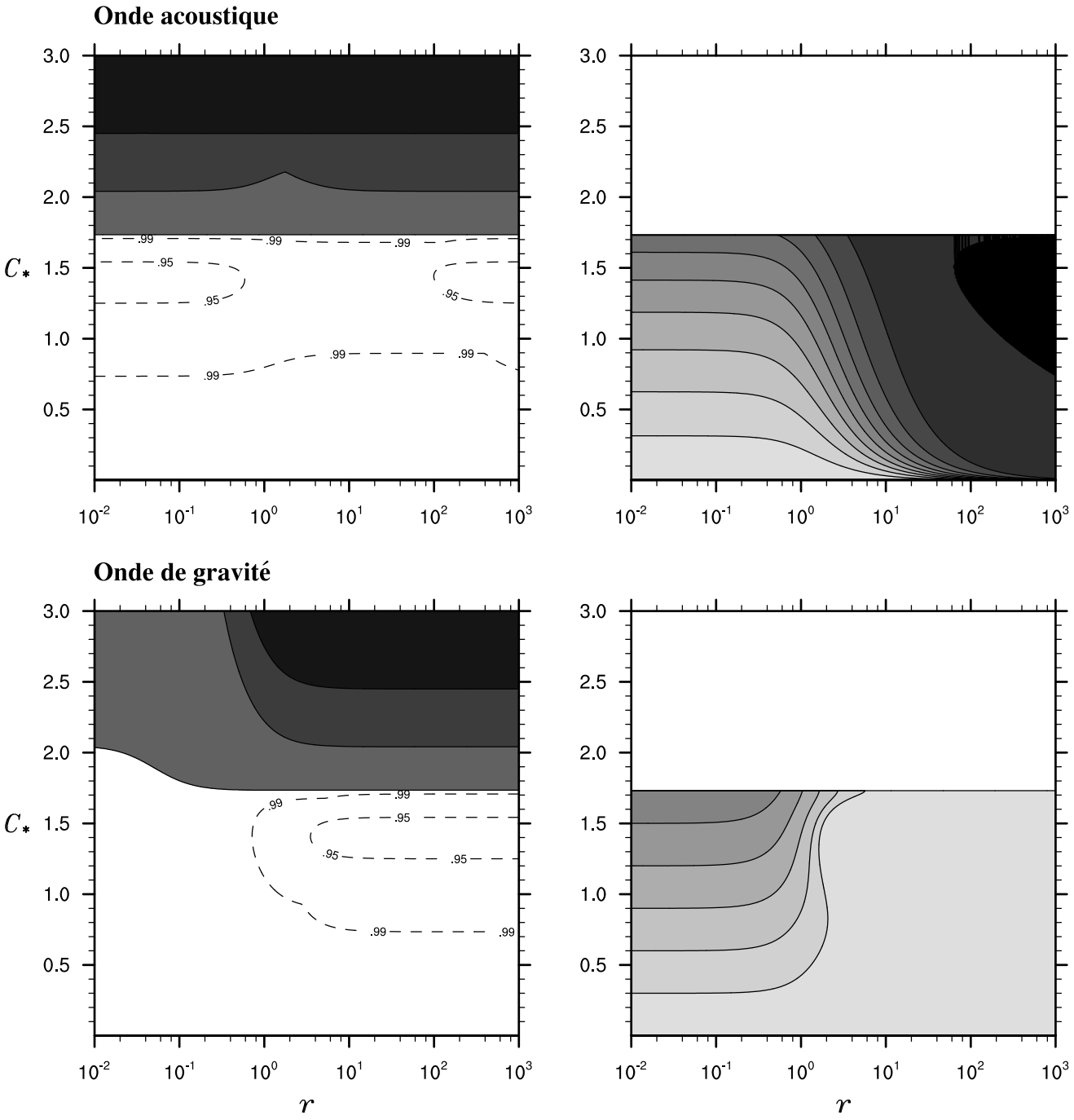


FIGURE III.5 – *Comportement des ondes modélisées par UJ3(1,3,2) UFpreF. À gauche, le maximum et le minimum des modules des valeurs propres correspondant à chacune des ondes, et à droite la phase de l'onde numérique.*

$C_* > 1{,}5$) et il semble qu'il existe une très large plage où la phase soit constante ($r > 1$ et $C_* > 1$), avec donc une vitesse de groupe nulle. Ceci est d'autant plus pénalisant que bien qu'une des ondes acoustiques soit fortement amortie dans cette zone, ce n'est pas le cas de la seconde.

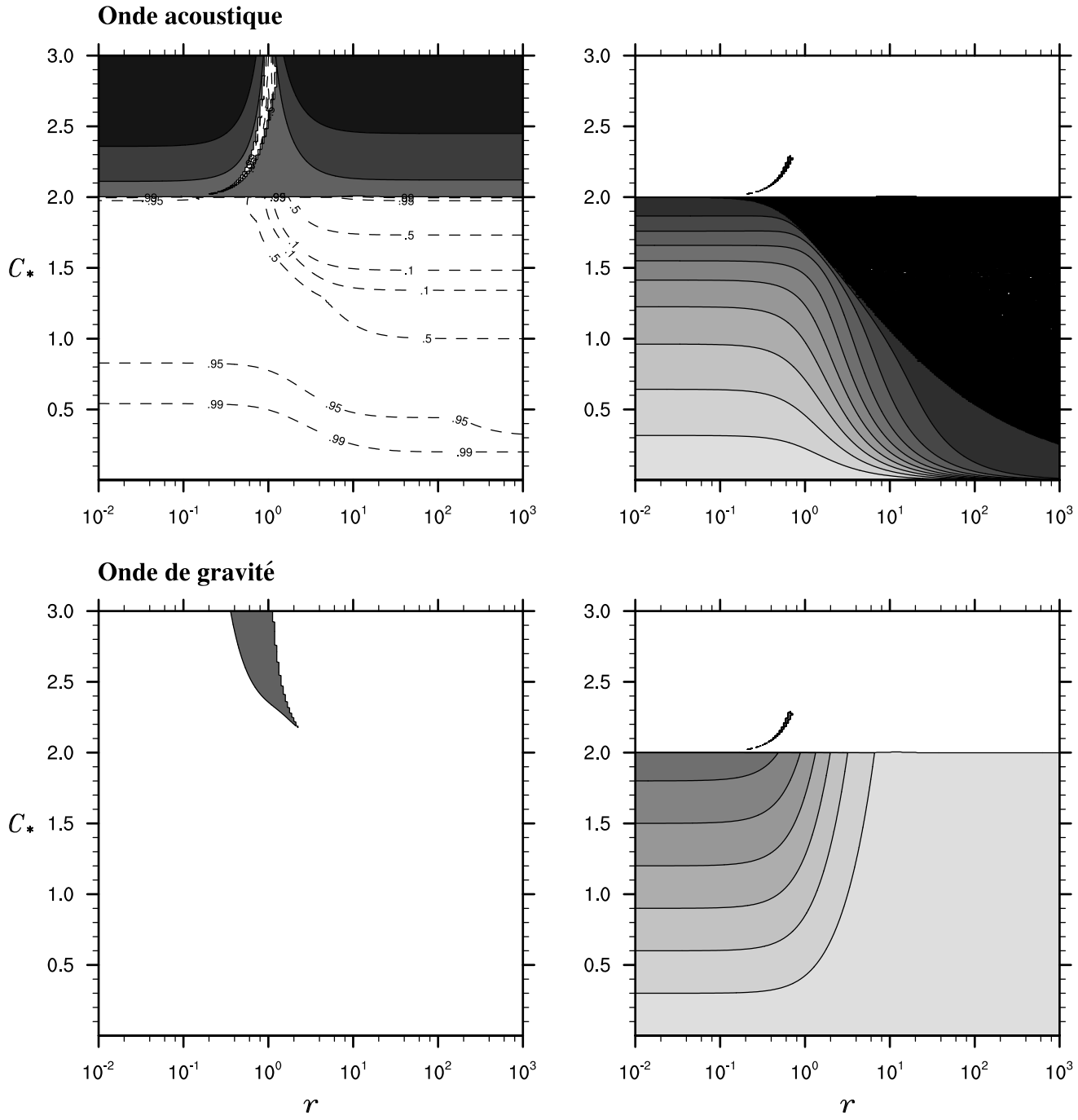


FIGURE III.6 – *Même graphe que Figure III.5, mais pour le schéma Trap2(2,3,2)(-1) UFpreF.*

La Figure III.8 montre la même étude que précédemment, mais pour la version Mixed. Il conserve la même stabilité que les versions UFpreF et UFpreB ($C_* \leq 2$), et présente deux avantages par rapport à ces schémas originaux. Le premier est de réduire l'amortissement des ondes aux grandes longueurs d'ondes horizontales. En revanche, ce schéma amortit les ondes horizontales de

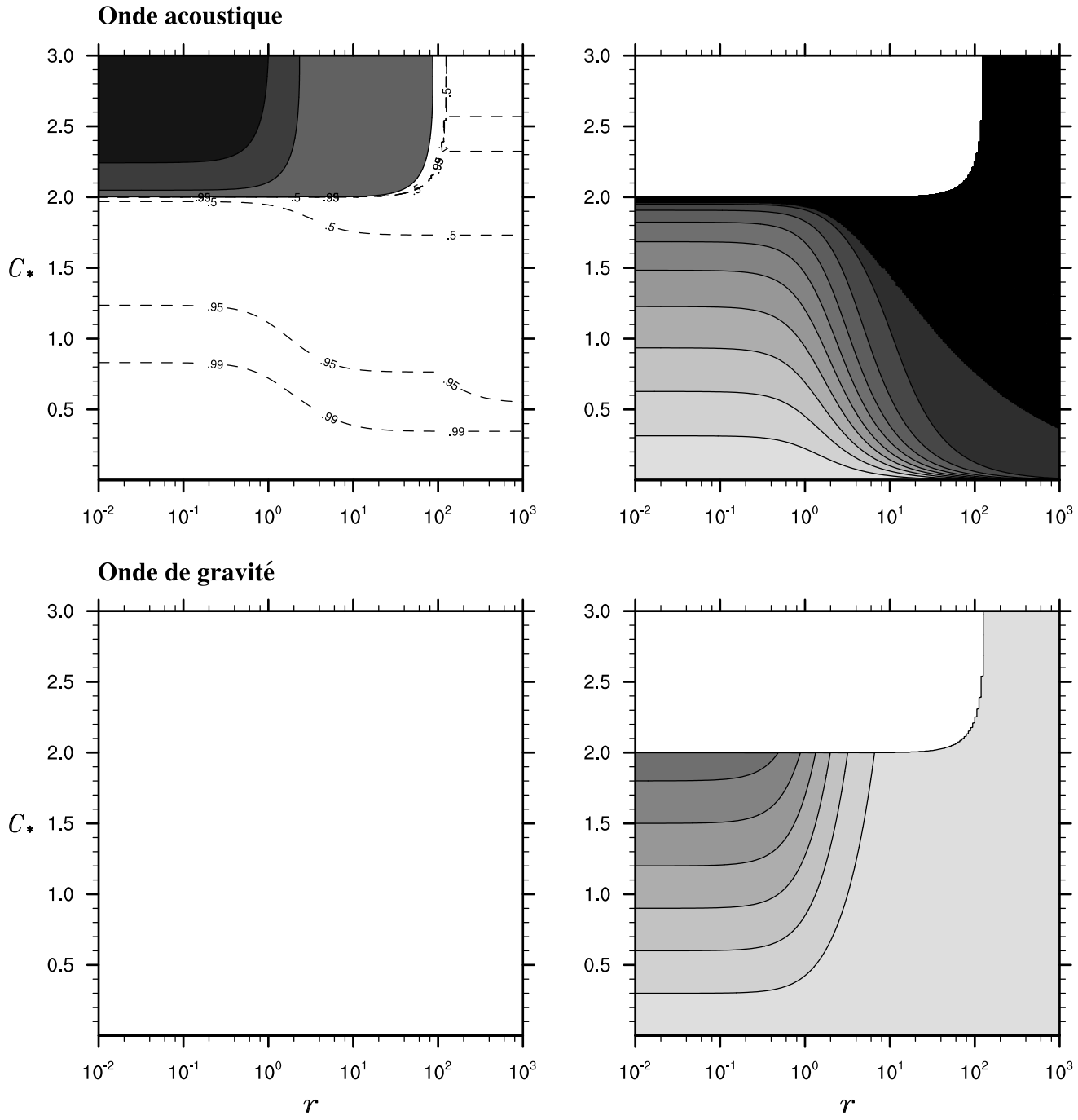


FIGURE III.7 – *Même graphe que Figure III.5, mais pour le schéma UJ3-Mixed.*

très faible longueur, ce qui, pour la PNT, n'est pas un grand inconvénient. Le second avantage se trouve dans le rétrécissement considérable de la zone où la vitesse de groupe est nulle. La meilleure représentativité des ondes acoustiques permet la diminution du risque de générer des instabilités dues à l'accumulation d'énergie.

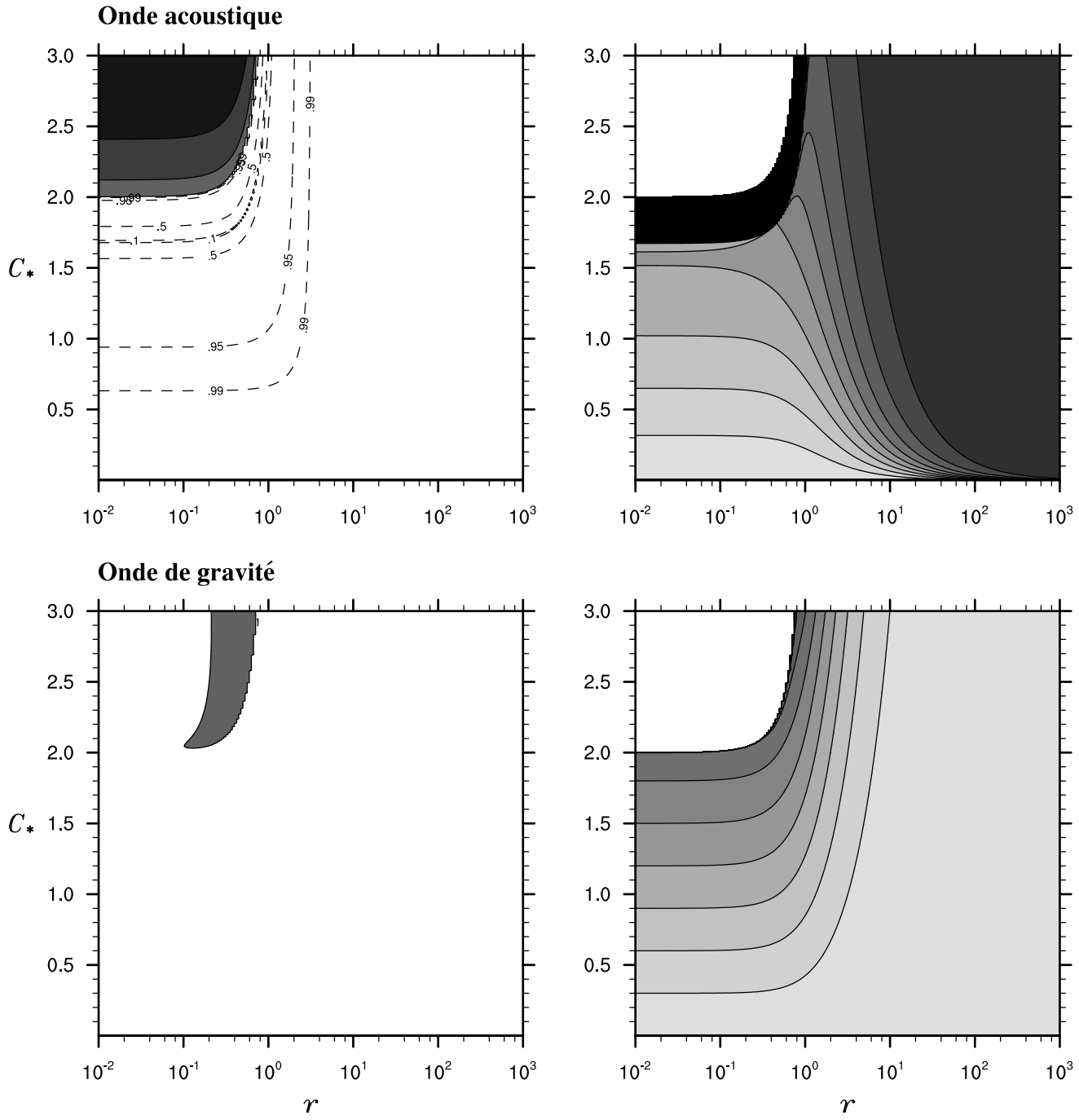


FIGURE III.8 – *Même graphe que Figure III.5, mais pour le schéma Trap2-Mixed.*

De manière générale pour ces deux schémas, il faut noter que la stabilité est limitée par le nombre de Courant acoustique horizontal, car ce sont les ondes acoustiques qui génèrent les instabilités. Néanmoins, il existe des cas de longueur d'ondes et des nombres de Courant pour lesquels les ondes de gravité sont instables. Nous avons tracé de tels cas dans la Figure III.4, où, en regar-

dant de près les valeurs propres associées aux ondes de gravité, il existe des cas où la norme est supérieure à 1.

La Figure III.7 montre les résultats de cette étude pour le UJ3-Mixed. Par rapport au traitement UFpreF, il y a un gain de stabilité dû au fait que maintenant il suffit que $C_* \leq 2$ (comme la version UFpreB). Il existe un relativement grand domaine d'annulation de la vitesse de groupe pour les ondes acoustiques pour $r > 10$ et $C_* > 1$. De plus, il semble que les ondes acoustiques soient amorties quelque soit le rapport d'aspect. Malgré ces petits défauts, il faut rappeler que ce schéma ne possède qu'une seule inversion sur la verticale, ce qui le rend intéressant car très économique.

Ces études de stabilité ont permis deux choses. La première, par la mise en lumière d'instabilités pathologiques en cas d'advection, nous avons pu éliminer certains des candidats présentés par Lock *et al.* [42] comme susceptibles d'être utilisés dans un contexte opérationnel. Ensuite, nous avons pu faire émerger un nouveau candidat, issu de Trap2(2,3,2)(-1), améliorant à la fois le comportement des ondes dans le cas où l'état de référence est au repos, mais aussi la stabilité en cas d'advection. L'ensemble des ces travaux d'analyses numériques sur la stabilité des schémas RK-IMEX HEVI les plus prometteurs a été récemment soumis au journal QJRMS pour publication. Le lecteur est invité à lire la version originale de cet article en Appendix A pour une discussion plus détaillée.

5 Discussion

Pour synthétiser ce chapitre, nous avons étendu les travaux de Lock *et al.* (2014) [42] sur deux aspects. Pour un système linéaire pleinement compressible plus complet que celui des ondes acoustiques 2D examinées, nous avons pu retrouver nombre de résultats similaires et noter que dans la plupart des cas, ces schémas ne déformaient pas les ondes de gravité (à l'exception du schéma UJ3(1,3,2) UFpreF). Le second aspect porte sur l'introduction d'un écoulement linéaire qui nous a permis de voir que nos schémas candidats initiaux devenaient, pour certains, fortement instables en présence d'advection. Pour palier ce problème, nous avons montré que l'introduction de deux tableaux de Butcher, non-nécessairement explicites, traitant les termes d'ajustements horizontaux permettaient de résoudre à la fois le problème de la stabilité, mais aussi pouvaient corriger certains comportements vis-à-vis des phases des ondes acoustiques. Nous avons de plus démontré que les nouveaux schémas que nous avons introduits étaient toujours d'ordre 2 en temps. Les deux nouveaux schémas issus de ces travaux, le UJ3-Mixed et le Trap2-Mixed, semblent devenir les candidats les plus prometteurs car il sont les plus stables jamais étudiée jusqu'alors, sans nécessiter plus de calculs que les versions originales. Le Trap2-Mixed possède peut-être un petit avantage par rapport à UJ3-Mixed du fait qu'il n'annule pas la vitesse de groupe pour les petites longueurs d'ondes verticales. Il faut néanmoins noter qu'il doit exister un schéma à quatre tableaux de Butcher construit en suivant le principe de ARK2(2,3,2) pour assurer la L-stabilité afin qu'il soit plus stable nos deux candidats. Pour cela, il faut étendre les travaux de Kennedy & Carpenter (2001) [33] qui fournit une méthodologie de construction de tels schémas respectant la condition de L-stabilité, mais uniquement pour trois tableaux de Butcher. Cette piste n'a pas été étudiée dans ces travaux.

Le prochain chapitre va donc s'intéresser à l'implémentation des schémas RK-IMEX que nous avons retenus pour le modèle non-linéaire. Pour mettre en lumière un autre avantage à utiliser ces versions Mixed, nous proposons de ne garder que deux schémas relativement stable mais fondamentalement différents : le Trap2(2,3,2)(-1) UFpreF qui est le plus stable des schémas à deux tableaux de Butcher pour cette version, et Trap2-Mixed. Néanmoins, les développement que nous allons réaliser reste valable pour n'importe quel schéma HEVI.

Chapitre IV

Traitement de l'orographie dans le cadre d'une approche HEVI

Les études de stabilité, qui ont été réalisées jusqu'à présent, n'ont pas tenu compte de l'impact de l'orographie sur la stabilité des schémas HEVI, du fait que le système linéaire $\mathscr{L}$ était supposé être sans orographie. Or, il est un fait avéré que les termes orographiques, associés à l'utilisation d'une coordonnée verticale épousant le relief, peuvent gravement nuire à la stabilité des schémas temporels (Ikawa (1988) [30]). En effet, l'utilisation d'une coordonnée verticale (telle que les coordonnées hybrides) fait apparaître les termes non-linéaires contenant la variabilité horizontale de l'orographie. Bénard (2005) [8] met en évidence que le traitement implicite de ces termes dans le schéma SI (à coefficients constants) via l'introduction d'une variable pronostique, dont la définition contient l'un de ces termes, permet d'obtenir un gain notable de stabilité. Pour le jeu de variables pronostiques considéré dans cette étude ($\mathbf{V}, w, T, q, \pi_s$), ces termes non-linéaires, notés X et Y, apparaissent respectivement dans la divergence 3D (et donc dans les équations de la température et de la déviation de pression non-hydrostatique) et dans les équations du mouvement horizontal.

Ce chapitre cherche à répondre aux trois questions suivantes : quel est l'impact de l'orographie sur la stabilité des schémas HEVI retenus ? Comment garantir la plus grande stabilité possibles de ces schémas en présence d'une orographie ? Enfin, comment appliquer les méthodes RK-IMEX, montrées comme les plus efficaces, au système complet d'Euler en coordonnée masse ?

Dans ce chapitre, afin de bien mettre en exergue l'impact de l'orographie sur les schémas et l'efficacité des solutions, nous proposons de réaliser ces analyses sur les deux schémas retenus du chapitre précédant, le Trap2(2,3,2)(-1) en version UFpreF et Mixed, qui sont, *a priori*, les plus comparables possibles. Il faut noter que certains résultats de ce chapitre s'appliquent au-delà de ce cadre. En effet, certains traitements peuvent être appliqués pour n'importe quel schéma HEVI et pour n'importe quelle coordonnée suivant le terrain.

1 Illustration du problème à l'aide du système acoustique bi-dimensionnel

Afin d'illustrer le problème lié aux termes croisés associés à la transformation pour une coordonnée verticale suivant le terrain, le système linéaire couplé décrivant la propagation des ondes acoustiques est ici considéré dans le plan vertical cartésien $(x-z)$. Ce système s'énonce simplement :

$$\partial_t u + \partial_x p = 0 \tag{IV.1}$$
$$\partial_t w + \partial_z p = 0 \tag{IV.2}$$
$$\partial_t p + \overline{c}_s^2 \left(\partial_x u + \partial_z w\right) = 0 \tag{IV.3}$$

où $\overline{c}_s$ est la vitesse du son. En effectuant un changement de coordonnée verticale en passant de la hauteur z à une coordonnée η quelconque telle que $\partial_z \eta$ soit strictement positif ou soit strictement négatif, les règles générales de la transformation de coordonnée s'écrivent :

$$\left(\frac{\partial \psi}{\partial \xi}\right)_z = \left(\frac{\partial \psi}{\partial \xi}\right)_\eta - \left(\frac{\partial z}{\partial \xi}\right)_\eta \left(\frac{\partial z}{\partial \eta}\right)^{-1} \left(\frac{\partial \psi}{\partial \eta}\right)_{x,t} \tag{IV.4}$$
$$\left(\frac{\partial \psi}{\partial z}\right)_{x,t} = \left(\frac{\partial z}{\partial \eta}\right)^{-1} \left(\frac{\partial \psi}{\partial \eta}\right)_{x,t} \tag{IV.5}$$

avec ψ une variable scalaire générique représentant l'une ou l'autre des variables u, w, ou p, et ξ une coordonnée x ou t. Lorsqu'une coordonnée est indicée au pied de la parenthèse cela signifie qu'elle est maintenue constante pendant que les autres coordonnées sont libres de varier. Par la suite, nous omettrons cette notation en n'oubliant pas que la dérivée partielle d'une variable par rapport à une coordonnée est opérée en supposant que les autres coordonnées du système sont fixées.

Considérons une coordonnée suivant le terrain de la forme $\eta = z - z_s$, où $z_s = z_s(x)$ est la hauteur du relief. Ce dernier est supposé avoir une pente uniforme telle que $z_s(x) = -sx$ avec s constante. Par conséquent, l'application des règles de transformation (IV.4) et (IV.5) éditées ci-dessus conduit à écrire le système en coordonnée (t, x, η) ainsi :

$$\partial_t u + \left(\partial_x p + s\,\partial_\eta p\right) = 0 \tag{IV.6}$$
$$\partial_t w + \partial_\eta p = 0 \tag{IV.7}$$
$$\partial_t p + \overline{c}_s^2 \left(\partial_x u + s\,\partial_\eta u\right) + \overline{c}_s^2\,\partial_\eta w = 0 \tag{IV.8}$$

Pour un domaine non-borné, le système (IV.6)-(IV.8) ci-dessus admet des solutions ondulatoires de la forme :

$$\hat{\psi} = \hat{\psi}_0 \exp[\hat{\imath}(k\,x + \ell\,\eta - \omega t)]$$

avec

$$\omega = \pm \overline{c}_s \sqrt{(k + s\ell)^2 + \ell^2}$$

où $(k, \ell) \in \mathbb{R}^2$ est le couple de nombres d'ondes horizontales et verticales. Rappelons que, par nature, ces solutions sont supposées maintenir une amplitude constante.

Considérons maintenant une discrétisation temporelle de type HEVI de ce système. Dans la mesure où les schémas RK-IMEX HEVI se sont révélés plus stables comparativement à l'approche forward/backward HEVI classique utilisée dans les méthodes au pas de temps fractionné, nous examinerons ici le cas d'une discrétisation Trap2(2,3,2)(-1) dans sa configuration UFpreF. De manière classique également, cette discrétisation implique une séparation des termes en une partie traitée explicitement notée $\mathscr{E}$ et une autre partie traitée implicitement notée $\mathscr{I}$, de sorte que :

$$\partial_t X = \mathscr{E}(X) + \mathscr{I}(X)$$

où X désigne le vecteur des variables pronostiques du système ici $X = {}^{\mathrm{t}}(u, w, p)$. Sous la condition UFpreF HEVI, cette séparation est effectuée de la façon suivante :

$$\mathscr{E} = \begin{bmatrix} 0 & 0 & -\partial_x + (\gamma - 1)s\partial_\eta \\ 0 & 0 & 0 \\ -\overline{c}_s^2\partial_x + \overline{c}_s^2(\gamma - 1)s\partial_\eta & 0 & 0 \end{bmatrix} \tag{IV.9}$$

$$\mathscr{I} = \begin{bmatrix} 0 & 0 & -\gamma\, s\partial_\eta \\ 0 & 0 & -\partial_\eta \\ -\overline{c}_s^2\gamma\, s\partial_\eta & -\overline{c}_s^2\partial_\eta & \end{bmatrix} \tag{IV.10}$$

où les termes horizontaux sont traités strictement explicitement, l'option Mixed développée dans le chapitre précédent n'étant pas envisagée dans cette illustration. Les termes d'ajustements verticaux (*ie :* $\partial_\eta w$ dans l'équation de la pression et $\partial_\eta p$ dans l'équation de la vitesse verticale) sont évalués implicitement. Les termes faisant intervenir la pente s sont traités soit explicitement (*ie :* comme ils le sont habituellement) si $\gamma = 0$ ou implicitement si $\gamma = 1$.

Le système est finalement discrétisé dans le temps via le schéma Trap2(2,3,2)(-1) UFpreF selon les étapes de calculs suivantes :

$$\frac{X^{(1)} - X^0}{\Delta t} = \mathscr{E}(X^0) + \mathscr{I}(X^0) \tag{IV.11}$$

$$\frac{X^{(2)} - X^0}{\Delta t} = \frac{1}{2}\left[\mathscr{E}(X^0) + \mathscr{E}(X^{(1)})\right] + \frac{1}{2}\left[\mathscr{I}(X^0) + \mathscr{I}(X^{(2)})\right] \tag{IV.12}$$

$$\frac{X^{+} - X^0}{\Delta t} = \frac{1}{2}\left[\mathscr{E}(X^0) + \mathscr{E}(X^{(2)})\right] + \frac{1}{2}\left[\mathscr{I}(X^0) + \mathscr{I}(X^{+})\right] \tag{IV.13}$$

avec, là encore, X^0 la valeur du vecteur des variables pronostiques au temps courant t et X^+ sa valeur au temps $t + \Delta t$. Les $X^{(1)}$, et $X^{(2)}$ sont des valeurs intermédiaires de calcul.

Stabilité

La stabilité de ce système discrétisé en temps est examinée en observant la même méthodologie que dans le chapitre précédent. L'analyse consiste donc à examiner le comportement des modes propres ondulatoires du système en remplaçant respectivement les opérateurs ∂_x et ∂_η par leurs

valeurs propres (dans l'espace des modes propres), $\hat{\imath}k$ et $\hat{\imath}\ell$. On obtient ainsi des matrices scalaires pour $\mathscr{E}$ et $\mathscr{I}$ définies par :

$$\mathbf{E} = -\hat{\imath}k\overline{c}_s \begin{bmatrix} 0 & 0 & (1+(1-\gamma)S_*)/\overline{c}_s \\ 0 & 0 & 0 \\ \overline{c}_s(1+(1-\gamma)S_*) & 0 & 0 \end{bmatrix} \tag{IV.14}$$

$$\mathbf{I} = -\hat{\imath}k\overline{c}_s \begin{bmatrix} 0 & 0 & \gamma S_*/\overline{c}_s \\ 0 & 0 & r \\ \overline{c}_s\gamma S_* & \overline{c}_s r & \end{bmatrix} \tag{IV.15}$$

avec $S_* = sr$, et $r = \ell/k$. Pour une résolution horizontale cible de l'ordre de 1 km les pentes peuvent aller jusqu'à 78% en valeur absolue (soit $s = \pm 0{,}78$), par exemple dans le massif des Alpes bernoises (Suisses) en Europe. Dans ce travail, nous nous limiterons à des rapports d'aspects positifs, ainsi qu'à des pentes s évoluant dans l'intervalle $[0, 1]$. De plus, puisque $r \in [10^{-2}; 10^3]$ (voir chapitre 1), le domaine de variations du paramètre S_* est pris dans l'intervalle $[0, 10^3]$.

Les matrices d'amplification intermédiaires $\mathbf{A}^{(j)}$ sont telles que $\mathbf{X}^{(j)} = \mathbf{A}^{(j)} \cdot \mathbf{X}^0$, pour $j \in [\![1; 2]\!]$, et la matrice d'amplification finale $\mathbf{A}$ est donnée par $\mathbf{X}^+ = \mathbf{A} \cdot \mathbf{X}^0$. En injectant ces formes dans (IV.11) and (IV.13), il vient :

$$\mathbf{A}^{(1)} = \mathbf{1} + \Delta t(\mathbf{E} + \mathbf{I}), \tag{IV.16}$$

$$\mathbf{A}^{(2)} = \left[\mathbf{1} - \tfrac{\Delta t}{2}\mathbf{I}\right]^{-1} \left[\mathbf{1} + \tfrac{\Delta t}{2}(\mathbf{E} + \mathbf{I}) + \tfrac{\Delta t}{2}\mathbf{E}\mathbf{A}^{(1)}\right], \tag{IV.17}$$

$$\mathbf{A} = \left[\mathbf{1} - \tfrac{\Delta t}{2}\mathbf{I}\right]^{-1} \left[\mathbf{1} + \tfrac{\Delta t}{2}(\mathbf{E} + \mathbf{I}) + \tfrac{\Delta t}{2}\mathbf{E}\mathbf{A}^{(2)}\right] \tag{IV.18}$$

Cas $\gamma = 0$ ***:***
Après quelques manipulations algébriques, la matrice d'amplification peut s'écrire en fonction du nombre de Courant ondulatoire horizontal $C_* = \overline{c}k\Delta t$, du rapport d'aspect r et du paramètre relié à la pente S_* de la manière suivante :

$$\mathbf{A} = \begin{bmatrix} \dfrac{(r^2 - 2(S_*+1)^2)\,C_*^2 + 4}{r^2C_*^2+4} & -\dfrac{2r(S_*+1)C_*^2}{r^2C_*^2+4} & \dfrac{\hat{\imath}(S_*+1)C_*((S_*+1)^2C_*^2 - 4)}{r^2C_*^2+4} \\ \dfrac{r(S_*+1)C_*^2((S_*+1)^2C_*^2 - 4)}{2r^2C_*^2+8} & \dfrac{r^2\left((S_*+1)^2C_*^2 - 2\right)C_*^2 + 8}{2r^2C_*^2+8} & \dfrac{\hat{\imath}rC_*((S_*+1)^2C_*^2 - 4)}{r^2C_*^2+4} \\ \dfrac{\hat{\imath}(S_*+1)C_*((S_*+1)^2C_*^2 - 4)}{r^2C_*^2 + 4} & \dfrac{\hat{\imath}\,rC_*((S_*+1)^2C_*^2 - 4)}{r^2C_*^2+4} & \dfrac{4 - (r^2 + 2(S_*+1)^2)\,C_*^2}{r^2C_*^2+4} \end{bmatrix} \tag{IV.19}$$

Cette matrice possède trois valeurs dont une trivialement égale à l'unité $\lambda_1 = 1$ associée au mode non-divergent. Les deux autres valeurs propres sont complexes conjuguées, notées communément λ_+ et λ_-, et sont liées aux deux modes acoustiques se propageant dans des directions opposées. Ces deux valeurs propres dépendent des trois paramètres C_*, r, et S_*. L'analyse se poursuit en se plaçant dans le cas asymptotique $r \to 0$, correspondant au cas limite des modes externes de Lamb, lesquelles se propagent exclusivement de manière horizontale. En procédant à l'analyse du module

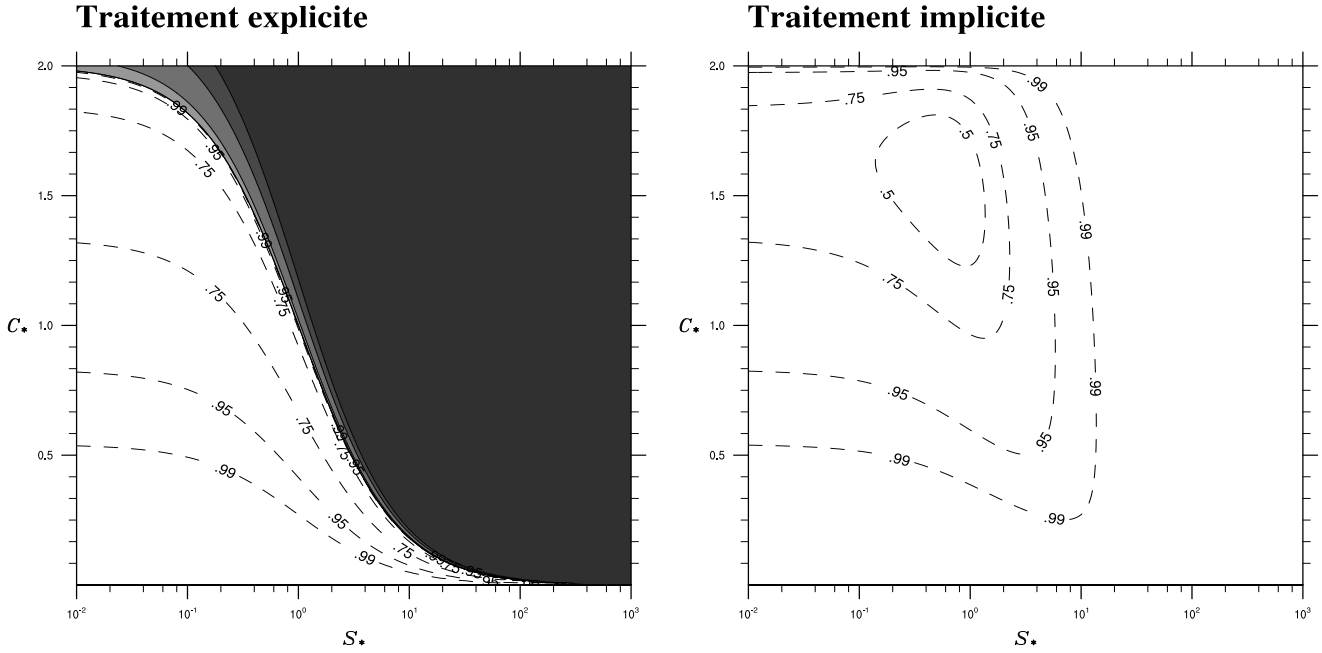


FIGURE IV.1 – *Stabilité du modes externes de Lamb en fonction de C_* et de S_*.*

des valeurs propres pour $r = 10^{-4}$, il apparaît qu'il existe une dépendance hyperbolique entre le nombre de Courant limite schéma $C_{\text{limite}*}$ et le paramètre S_*, qui est illustrée à la Figure (IV.1). Ce comportement peut également être vérifié en calculant analytiquement $\Gamma_\pm^2 = |\lambda_\pm|^2$ le carré du module des valeurs propres de la matrice d'amplification donnée par (IV.19), puis en effectuant un développement limité à l'ordre 2 en r, nous obtenons :

$$\Gamma_\pm^2 = 1 + \frac{C_*^4}{16}\left(1+S_*\right)^4\left[C_*^2\left(1+S_*\right)^2 - 4\right] + O(r^4) \tag{IV.20}$$

Le schéma est donc stable pour les modes acoustiques externes à condition que $C_*\left(1+S_*\right)p \leq 2$.

Nous constatons que le traitement explicite des termes associés à la transformation en coordonnée suivant le relief impose une contrainte sur le nombre de Courant horizontal. Il est important de remarquer que la sévérité de cette contrainte dépend du signe paramètre S_* et de sa position par rapport à 1. Autrement dit, cette analyse très simplifiée semble indiquer que la stabilité d'un mode ondulatoire défini par les nombres d'ondes (k,ℓ) proches du mode externe $\ell \ll k$ dépend de l'angle d'incidence de mode ondulatoire par rapport à la pente, et du signe de la pente.

Cas $\gamma = 1$ ***:***
Ce cas correspond au traitement implicite des termes relatifs à la présence du relief dans le système. La matrice d'amplification $\mathbf{A}$ s'obtient en résolvant (IV.18), mais cette fois-ci la matrice à inverser $[\mathbf{1} - (\Delta t/2)\mathbf{I}]$ est un peu plus compliquée car elle contient les termes additionnels faisant intervenir S_*. Le résultat de ce calcul est bien trop long pour être présenté de manière succincte dans ce chapitre. Néanmoins en procédant comme pour le cas $\gamma = 0$, (*ie :* en faisant tendre le rapport d'aspect r vers zéro) il est possible d'exhiber le taux de croissance maximum du schéma de manière analytique. Ainsi l'impact du schéma temporel sur l'amplitude des modes acoustiques externes est déterminé en première approximation pour de faibles valeurs de r^2, par :

$$\Gamma_{\pm}^2 = 1 - \frac{C_*^4 \left(1 + S_*\right)^2}{(4 + S_*^2 C_*^2)^2} \left(4 - C_*^2\right) + O(r^4) \tag{IV.21}$$

Cela implique donc que le traitement implicite des termes croisés issus de la coordonnée épousant le relief à un effet neutre, voir amortissant, sur les modes se propageant de manière quasi-horizontale, et, qui plus est, sur les modes acoustiques externes, si la contrainte CFL $C_* \leq 2$ est respectée. Cette analyse est corroborée par la Figure IV.1 mettant en évidence l'amortissement des modes tels que $r = 10^{-4}$ pour un traitement implicite de ces termes croisés en fonction de S_* et de C_*.

2 Application au système d'Euler en coordonnée masse

Afin d'étendre l'étude précédente pour le système linéarisé d'Euler, il est d'abord nécessaire de mettre en évidence les termes orographiques. Pour cela, nous pouvons séparer les termes X et Y en deux contributions :

$$\mathsf{X} = \overbrace{\frac{1}{RT}\nabla\phi_s \cdot \frac{\pi}{m}\partial_\eta \mathbf{V}}^{\mathsf{X}_s} + \delta\mathsf{X} \tag{IV.22}$$

$$\mathsf{Y} = \underbrace{\nabla\phi_s \frac{\pi}{m}\partial_\eta(\mathrm{e}^q - 1)}_{\mathsf{Y}_s} + \delta\mathsf{Y} \tag{IV.23}$$

avec :

$$\delta\mathsf{X} = \frac{1}{RT}\nabla(\phi - \phi_s) \cdot \frac{\pi}{m}\partial_\eta \mathbf{V} \tag{IV.24}$$

$$\delta\mathsf{Y} = \nabla(\phi - \phi_s)\frac{\pi}{m}\partial_\eta(\mathrm{e}^q - 1) \tag{IV.25}$$

La démarche, semblable à la section précédente, consiste en un traitement implicite des termes X_s et Y_s. Dans le cas du système discrétisé, cela requiert une réflexion sur les algorithmes à utiliser, et les opérateurs à créer. Mais avant de proposer un traitement spécifique pour ces termes, il convient de déterminer comment traiter les autres termes du système.

Le chapitre précédent montre qu'il est pertinent d'utiliser au moins quatre schémas Runge-Kutta pour intégrer le système d'Euler. Le traitement des termes advectifs nécessitent l'utilisation d'un schéma explicite. La contrainte HEVI impose qu'au moins un des schémas (celui intégrant

les termes d'ajustements verticaux) possède, au moins, une itération implicite. Dans les systèmes des précédents chapitres, le reste des termes (le gradient de la pression et la divergence horizontale du vent) pouvait être intégré par des schémas tels qu'ils ne soient pas tous deux implicites pour la même itération (pour ne pas avoir à inverser un problème sur l'horizontale). Ces deux schémas sont donc non-nécessairement explicites. De plus, une étude au deuxième chapitre montre que le terme d'auto-convection devait être traité au même instant que la divergence horizontale du vent, afin d'assurer que l'équation de structure discrète soit analogue à l'équation de structure continue. Enfin, les résidus $\delta\mathsf{X}$ et $\delta\mathsf{Y}$ n'étant ni linéaires, ni directement reliés à l'orographie, ils peuvent donc être traités de manière explicite comme l'advection. Pour résumer ces remarques, et en reprenant les notations du chapitre précédent, les équations suivantes présentent ce partitionnement des processus :

$$\begin{aligned}
\frac{d}{dt}\mathbf{V} + \overbrace{RT\left(\frac{\nabla\pi}{\pi} + \nabla q\right) + \nabla\phi}^{\mathscr{P}} + \overbrace{\delta\mathsf{Y}}^{\mathscr{E}'} + \overbrace{\mathsf{Y}_s}^{\mathscr{Y}} &= 0 \\
\frac{d}{dt}w - \underbrace{\frac{g}{m}\left(\partial_\eta\pi(\mathrm{e}^q - 1)\right)}_{\mathscr{I}'} &= 0 \\
\frac{d}{dt}T + \frac{RT}{C_v}\left(\underbrace{\nabla\cdot\mathbf{V}}_{\mathscr{U}} - \underbrace{\frac{\mathrm{e}^q}{H}\frac{\pi}{m}\partial_\eta w}_{\mathscr{I}'} + \underbrace{\delta\mathsf{X}}_{\mathscr{E}'} + \underbrace{\mathsf{X}_s}_{\mathscr{X}}\right) &= 0 \\
\frac{d}{dt}q + \underbrace{\frac{\dot\pi}{\pi}}_{\mathscr{U}} + \frac{C_p}{C_v}\left(\underbrace{\nabla\cdot\mathbf{V}}_{\mathscr{U}} - \underbrace{\frac{\mathrm{e}^q}{H}\frac{\pi}{m}\partial_\eta w}_{\mathscr{I}'} + \underbrace{\delta\mathsf{X}}_{\mathscr{E}'} + \underbrace{\mathsf{X}_s}_{\mathscr{X}}\right) &= 0 \\
\partial_t\pi_s + \underbrace{\int_0^1 \nabla\cdot(m\mathbf{V})\mathrm{d}\eta}_{\mathscr{U}} &= 0
\end{aligned}$$

avec les termes advectifs qui sont traités par $\mathscr{E}'$

Ainsi, par rapport au système linéaire, s'adjoignent deux parties supplémentaires, qui concentrent les termes issus de la présence de l'orographie. Il est possible d'intégrer ces termes à l'aide de deux tableaux de Butcher supplémentaires dont, pour le moment, aucune hypothèse sur la nature explicite ou implicite n'est établie. Il faut néanmoins que l'ensemble de ces matrices respectent les

conditions d'ordre deux définies par (III.8)-(III.10). Ainsi, le schéma s'écrit :

$$\begin{aligned}\frac{X^{(j)}-X^0}{\Delta t} =& \sum_{i=1}^{j-1}\tilde{a}_{ij}\mathscr{E}'(X^{(i)}+\sum_{i=1}^{j}a_{ij}\mathscr{I}'(X^{(i)}) \\ &+\sum_{i=1}^{j}\left[a_{ij}^u\mathscr{U}(X^{(i)})+a_{ij}^p\mathscr{P}(X^{(i)})\right]+\sum_{i=1}^{j}\left[a_{ij}^x\mathscr{X}(X^{(i)})+a_{ij}^y\mathscr{P}(Y^{(i)})\right] \qquad \text{(IV.26)}\\ \frac{X^{+}-X^0}{\Delta t} =& \sum_{j=1}^{\nu}\tilde{b}_{j}\mathscr{E}'(X^{(j)})+\sum_{i=1}^{j}b_{j}\mathscr{I}'(X^{(j)}) \\ &+\sum_{j=1}^{\nu}\left[b_{j}^u\mathscr{U}(X^{(j)})+b_{j}^p\mathscr{P}(X^{(j)})\right]+\sum_{j=1}^{\nu}\left[b_{j}^x\mathscr{X}(X^{(j)})+b_{j}^y\mathscr{Y}(X^{(j)})\right] \qquad \text{(IV.27)}\end{aligned}$$

où $\{c^x,\mathcal{A}^x,b^x\}$ et $\{c^y,\mathcal{A}^y,b^y\}$ sont les tableaux de Butcher résumant les deux schémas Runge-Kutta réalisant l'intégration temporelle respectivement des processus $\mathscr{X}$ et $\mathscr{Y}$.

La grande différence entre la résolution d'un système linéaire et non-linéaire réside dans l'inversion du problème implicite. En effet, dans le cas d'un système linéaire, il suffit de résoudre un problème matriciel pour lequel de nombreuses méthodes peuvent être utilisées. Dans le cas non-linéaire, à chaque étape $j\in[\![1;\nu]\!]$, le problème inverse s'écrit :

$$\mathbf{V}^{(j)}+a_{jj}^p\Delta t RT^{(j)}\left(\frac{\nabla\pi^{(j)}}{\pi^{(j)}}+\nabla q^{(j)}\right)+a_{jj}^p\Delta t\nabla\phi^{(j)}+a_{jj}^y\Delta t\Upsilon^{(j)}=\mathbf{V}^{\bullet} \qquad \text{(IV.28)}$$

$$w^{(j)}-ga_{jj}\Delta t(\tilde{\partial}+\mathrm{I})(\mathrm{e}^{q^{(j)}}-1)=w^{\bullet} \qquad \text{(IV.29)}$$

$$T^{(j)}+\Delta t\frac{RT^{(j)}}{C_v}\left(a_{jj}^u\nabla\cdot\mathbf{V}^{(j)}-a_{jj}\frac{\mathrm{e}^{q^{(j)}}}{H^{(j)}}\tilde{\partial}w^{(j)}+a_{jj}^x\mathrm{e}^{q^{(j)}}\mathsf{X}^{(j)}\right)=T^{\bullet} \qquad \text{(IV.30)}$$

$$q^{(j)}+a_{jj}^u\Delta t\left(\frac{\dot{\pi}}{\pi}\right)^{(j)}+\Delta t\frac{C_p}{C_v}\left(a_{jj}^u\nabla\cdot\mathbf{V}^{(j)}-a_{jj}\frac{\mathrm{e}^{q^{(j)}}}{H^{(j)}}\tilde{\partial}w^{(j)}+a_{jj}^x\mathrm{e}^{q^{(j)}}\mathsf{X}^{(j)}\right)=q^{\bullet} \qquad \text{(IV.31)}$$

$$\pi_s^{(j)}+a_{jj}^u\Delta t\int_0^1\nabla\cdot(m^{(j)}\mathbf{V}^{(j)})\mathrm{d}\eta=\pi_s^{\bullet} \qquad \text{(IV.32)}$$

avec :

$$\begin{aligned}X^{\bullet} =& X^0+\Delta t\sum_{i=1}^{j-1}\tilde{a}_{ij}\mathscr{E}'(X^{(i)})+\Delta t\sum_{i=1}^{j-1}a_{ij}\mathscr{I}'(X^{(i)}) \\ &+\Delta t\sum_{i=1}^{j-1}\left[a_{ij}^u\mathscr{U}(X^{(i)})+a_{ij}^p\mathscr{P}(X^{(i)})\right]+\Delta t\sum_{i=1}^{j-1}\left[a_{ij}^x\mathscr{X}(X^{(i)})+a_{ij}^y\mathscr{Y}(X^{(i)})\right] \qquad \text{(IV.33)}\end{aligned}$$

Rappelons que la contrainte HEVI impose que les coefficients a_{jj}^u et a_{jj}^p ne soient pas simultanément non-nuls. En effet, dans le cas contraire, il est aisé de voir que cela implique la résolution d'un problème implicite pour la direction horizontale. De plus, afin de maintenir la meilleure stabilité possible pour une orographie quelconque, il est possible d'envisager un traitement implicite des

termes $\mathscr{X}_s$ et $\mathscr{Y}_s$ (*ie :* a^x_{jj} et/ou a^y_{jj} non-nul). Le système ci-avant peut se résumer par :

$$f_j(X^{(j)}) = X^\bullet \tag{IV.34}$$

Approche itérative quasi-Newton

Comme l'opérateur f_j est non-linéaire, il est très difficile de résoudre l'équation (IV.34). L'une des méthodes consiste à approcher f_j par une application linéaire $f_j^\star$. Cet opérateur est dépendant de l'état autour duquel il est calculé. Pour minimiser la distance entre cet état et $X^{(j)}$, et ainsi favoriser la convergence rapide de ces itérations, il est préférable de choisir $X^{(j-1)}$ (et X^0 pour l'initialisation). Ainsi, la linéarisation implique :

$$f_j(X^{(j)}) \approx f_j(X^{(j-1)}) + f_j^\star(X^{(j-1)}) \cdot (X^{(j)} - X^{(j-1)})$$

En faisant des itérations successives $X^{k=0} = X^{(j-1)}$ et $X^{(j)} = X^{k=N_{iter}}$, le système (IV.34) se résout :

$$\begin{aligned} f_j^\star(X^k) \cdot X^{k+1} &= X^\bullet - f_j(X^k) + f_j^\star(X^k) \cdot X^k \\ &= X^{\bullet,k} \end{aligned} \tag{IV.35}$$

Cette méthode quasi-Newton converge d'autant plus vite que l'opérateur $f_j^\star$ est proche du système linéaire tangent à l'opérateur f. Nous allons définir $f_j^\star$ à partir de la linéarisation du terme suivant :

$$e^{q^{k+1}} - 1 \approx e^{q^k}(1 + q^{k+1} - q^k) - 1$$

De plus, nous imposons que les coefficients devant les opérateurs de dérivations soient explicites. Ce traitement assure que les ondes acoustiques soient traitées implicitement. Enfin, nous assurons que la direction horizontale soit évaluée explicitement en imposant que certains termes soient évalués à l'itération k au lieu de $k+1$. Ainsi, le quasi-Newton que nous résolvons à chaque itération est :

$$\begin{aligned} \mathbf{V}^{k+1} + a^y_{jj}\Delta t \Upsilon^{k+1} =& \mathbf{V}^{\bullet,k} - a^p_{jj}\Delta t \nabla \phi^k \\ & - a^p_{jj}\Delta t R T^k \left(\frac{\nabla \pi^k}{\pi^k} + \nabla q^k \right) \end{aligned} \tag{IV.36}$$

$$w^{k+1} - g a_{jj}\Delta t(\tilde{\partial} + \mathrm{I})(\mathrm{e}^{q^k} q^{k+1}) = w^{\bullet,k} \tag{IV.37}$$

$$T^{k+1} + \Delta t \frac{RT^k e^{q^k}}{C_v} \left(-\frac{a_{jj}}{H^k}\tilde{\partial} w^{k+1} + a^x_{jj} \mathsf{X}^{k+1} \right) = T^\bullet - a^u_{jj}\Delta t \frac{RT^k}{C_v} \nabla \cdot \mathbf{V}^k \tag{IV.38}$$

$$\begin{aligned} q^{k+1} + \Delta t \frac{C_p}{C_v} e^{q^k} \left(-\frac{a_{jj}}{H^k}\tilde{\partial} w^{k+1} + a^x_{jj} \mathsf{X}^{k+1} \right) =& q^\bullet - a^u_{jj}\Delta t \left(\frac{\dot{\pi}}{\pi} \right)^k \\ & - a^u_{jj}\Delta t \frac{C_p}{C_v} \nabla \cdot \mathbf{V}^k \end{aligned} \tag{IV.39}$$

$$\pi_s^{k+1} = \pi_s^\bullet - a^u_{jj}\Delta t \int_0^1 \nabla \cdot (m^k \mathbf{V}^k) \mathrm{d}\eta \tag{IV.40}$$

avec $H^k = RT^k/g$ et :

$$\mathsf{Y}^{k+1} = \nabla\phi_s(\tilde{\partial} + \mathrm{I})(\mathrm{e}^{q^k} q^{k+1})$$
$$\mathsf{X}^{k+1} = \frac{\nabla\phi_s}{g} \cdot \tilde{\partial}\mathbf{V}^{k+1}$$

La forme du problème à résoudre soulève une remarque sur la géométrie du domaine. En effet, dans ce système, les opérateurs verticaux sont définis par rapport à la pression de surface π_s. Ainsi, dès que a^u_{jj} est non-nul, alors la géométrie verticale est à mettre à jour à chaque itération du quasi-Newton.

Pour un modèle opérationnel, la résolution de ce système s'effectue dans l'espace spatialement discrétisé, de sorte qu'il s'exprime comme une équation matricielle. Une technique consiste à calculer l'inverse de la matrice. Ce procédé demeure à la fois coûteux et peu précis car, *a priori*, cette matrice n'a aucune propriété permettant une inversion exacte. Une autre option alors propose une technique d'inversion par substitution. Celle-ci nécessite de définir la composée des opérateurs verticaux. Cette discussion seule pourra déterminer la faisabilité d'un traitement implicite simultanée des termes orographiques. Pour cette réaliser cette résolution implicite, nous allons présenter les opérateurs spatiaux discrétisés.

Définition des opérateurs spatiaux discrets

Le calcul des dérivations horizontales s'effectue dans l'espace spectral. Ce choix est motivé par deux raisons. La première permet de s'affranchir des contraintes de Mahrer (1984) [43] qui montre que, dans le cas d'une discrétisation horizontale de type différences finies, pour obtenir une représentation fidèle du relief, il est nécessaire que le rapport des mailles vérifie certaines relations. De plus, comme il est ici question de stabilité, il convient d'éviter toutes discussions sur l'origine de comportements potentiellement imputables à la discrétisation horizontale pour mesurer au mieux les différents comportements entre les schémas temporels.

En revanche, les opérateurs verticaux sont discrétisés par des différences finies. Le domaine est partitionné en L niveaux définissant la grille de Lorentz. Cela signifie que les variables $\mathbf{V}$, T et q sont définies au centre des mailles, alors que w et π sont définis sur les interfaces, afin d'assurer une précision d'ordre deux. Pour l compris en 1 et L, on note X_l la variable X prise au niveau l et $X_{\tilde{l}}$ sa valeur à l'interface du niveau l et $l+1$. De plus, $X_{\tilde{0}} = X_T$ est la valeur au sommet du modèle de la variable X. La Figure IV.2 illustre ces dispositions. Les propriétés sur les opérateurs intégraux continus ($\mathcal{C}_1$)-($\mathcal{C}_2$) doivent également être vérifiées par les opérateurs discrets. Budnovà *et al.* (1995) [9] décrivent l'ensemble des opérations menant à ces définitions. *In fine*, la géométrie du problème s'exprime à l'aide des termes métriques suivants :

$$\pi_{\tilde{l}} = A_{\tilde{l}} + B_{\tilde{l}}\pi_s$$
$$\delta\pi_l = \pi_{\tilde{l}} - \pi_{\tilde{l}-1}$$
$$\pi_l = \sqrt{\pi_{\tilde{l}-1}\pi_{\tilde{l}}}$$
$$\delta_l = \delta\pi_l/\pi_l$$
$$\alpha_l = 1 - \sqrt{\pi_{\tilde{l}-1}/\pi_{\tilde{l}}}$$

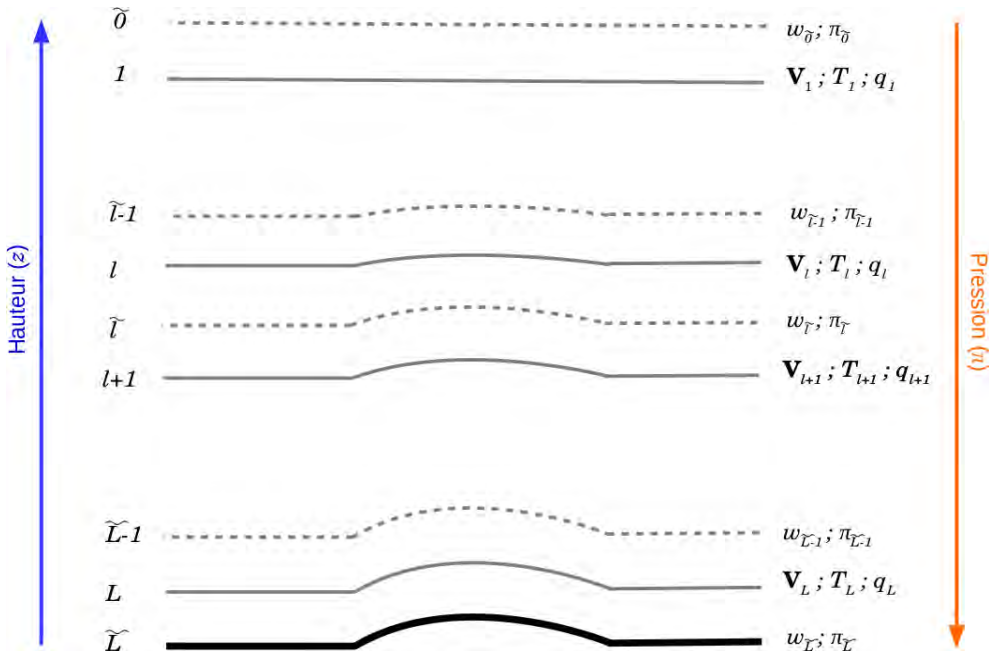


FIGURE IV.2 – *Illustration des niveaux verticaux du modèle.*

avec des valeurs spéciales au sommet :

$$\begin{aligned}\delta_1 &= 1 + C_p/R \\ \pi_1 &= \delta\pi_1/\delta_1 \\ \alpha_1 &= 1\end{aligned}$$

Grâce à ces termes, les opérateurs intégraux se définissent par :

$$\{\mathbf{G}X_l\}_l = \alpha_l X_l + \sum_{k=l+1}^{L} \delta_k X_k$$

$$\{\mathbf{S}X_l\}_l = \alpha_l X_l + \frac{1}{\pi_l}\sum_{k=1}^{l-1} \delta\pi_k X_k$$

$$\{\mathbf{N}X_l\}_l = \sum_{k=1}^{L} \delta\pi_k X_k$$

Les opérateurs dérivés sont eux définis par :

$$D_{\tilde{l}}^1 X_l \equiv \left\{\left(\tilde{\partial} + \mathrm{I}\right) X_l\right\}_{\tilde{l}} = \frac{\pi_{l+1}X_{l+1} - \pi_l X_l}{\pi_{l+1} - \pi_l}$$

$$D_l^0 X_{\tilde{l}} \equiv \left\{\tilde{\partial} X_{\tilde{l}}\right\}_l = \frac{X_{\tilde{l}} - X_{\tilde{l}-1}}{\delta_l}$$

Le Laplacien vertical $\mathbf{L}_v$ se définit en combinant les deux opérateurs D^0 et D^1 :

$$\{\mathbf{L}_v X_l\}_l \equiv \left\{\tilde{\partial}\left(\tilde{\partial} + \mathrm{I}\right) X_l\right\}_l = A_l X_{l-1} + B_l X_l + C_l X_{l+1}$$

où :

$$A_l = \frac{1}{\delta_l}\frac{\pi_{l-1}}{\pi_l - \pi_{l-1}}$$
$$C_l = \frac{1}{\delta_l}\frac{\pi_{l+1}}{\pi_{l+1} - \pi_l}$$
$$B_l = -(A_l + C_l)$$

et les conditions spéciales à la surface :

$$A_L = \frac{1}{\delta_L}\frac{\pi_{L-1}}{\pi_L - \pi_{L-1}}$$
$$B_L = -\frac{1}{\delta_L}\frac{\pi_L}{\pi_L - \pi_{L-1}}$$

L'utilisation de la coordonnée masse fait apparaître les termes X et Y dans le système. Ces variables sont particulières car elles nécessitent la définition d'opérateurs de dérivations verticales dont l'espace de départ et l'espace d'arrivée sont les niveaux l. Les valeurs sur les interfaces des variables définies sur les niveaux sont toutes calculées par une interpolation entre les niveaux l et $l+1$ donnée par :

$$X_{\tilde{l}} = \epsilon_l X_l + (1-\epsilon_l) X_{l+1}$$

avec :

$$\epsilon_l = \frac{\delta_{l+1} - \alpha_{l+1}}{\delta_{l+1} - \alpha_{l+1} + \alpha_l}$$

En utilisant ces définitions, les opérateurs des termes croisés X et Y sont donnés par :

$$D_l^1(\mathrm{e}^q q) \equiv \left\{\left(\tilde{\partial} + \mathrm{I}\right)\mathrm{e}^{q_l} q_l\right\}_l = \frac{\pi_{\tilde{l}}\mathrm{e}^{q_{\tilde{l}}}}{\delta\pi_l}(1-\epsilon_l)q_{l+1} + \frac{\pi_{\tilde{l}}\mathrm{e}^{q_{\tilde{l}}}\epsilon_l - \pi_{\tilde{l}-1}\mathrm{e}^{q_{\tilde{l}-1}}(1-\epsilon_{l-1})}{\delta\pi_l} q_l$$
$$- \frac{\pi_{\tilde{l}-1}\mathrm{e}^{q_{\tilde{l}-1}}}{\delta\pi_l}\epsilon_{l-1}q_{l-1} \qquad \text{(IV.41)}$$

$$D_l^0 \mathbf{V}_l \equiv \left\{\tilde{\partial}\mathbf{V}_{\tilde{l}}\right\}_l = \frac{(1-\epsilon_l)\mathbf{V}_{l+1} - (1-\epsilon_l-\epsilon_{l-1})\mathbf{V}_l - \epsilon_{l-1}\mathbf{V}_{l-1}}{\delta_l} \qquad \text{(IV.42)}$$

Pour l'ensemble de ces opérateurs, il est nécessaire d'imposer des conditions aux bords afin d'être en mesure de pouvoir calculer ces dérivées. Pour $\mathbf{V}$, nous imposons une condition de glissement en haut et à la surface du domaine (ce qui s'apparente à des conditions de Neumann), et pour les conditions sur q, nous imposons également un glissement à la surface, et au sommet, une condition élastique (condition de Dirichlet [1]) :

$$\mathbf{V}_{\tilde{0}} = \mathbf{V}_1, \quad \mathbf{V}_{\tilde{L}} = \mathbf{V}_L \qquad \text{(IV.43)}$$
$$q_{\tilde{0}} = 0, \quad q_{\tilde{L}} = q_L \qquad \text{(IV.44)}$$

1. Johann Peter Gustav Lejeune Dirichlet (1805-1859) : mathématicien allemand

Ainsi, aux bords, les opérateurs précédents sont définis par :

$$D_{\tilde{1}}^1(\mathrm{e}^q q)_l = \frac{\pi_{\tilde{1}}\mathrm{e}^{q_{\tilde{1}}}}{\delta\pi_1}(\epsilon_1 q_1 + (1-\epsilon_1)q_2) \tag{IV.45}$$

$$D_{\tilde{L}}^1(\mathrm{e}^q q)_l = \frac{\pi_{\tilde{L}}\mathrm{e}^{q_{\tilde{L}}} + (\epsilon_{L-1}-1)\pi_{\tilde{L}-1}\mathrm{e}^{q_{\tilde{L}-1}}}{\delta\pi_L} q_L - \frac{\pi_{\tilde{L}-1}\mathrm{e}^{q_{\tilde{L}-1}}\epsilon_{L-1}}{\delta\pi_L} q_{L-1} \tag{IV.46}$$

$$D_1^0 \mathbf{V}_l = \frac{1-\epsilon_1}{\delta_1}(\mathbf{V}_2 - \mathbf{V}_1) \tag{IV.47}$$

$$D_L^0 \mathbf{V}_l = \frac{\epsilon_{L-1}}{\delta_L}(\mathbf{V}_L - \mathbf{V}_{L-1}) \tag{IV.48}$$

Grâce à ces opérateurs discrets, les variables X_s et Y_s s'écrivent facilement :

$$\mathsf{X}_{sl} = \frac{\nabla\phi_s}{RT_l} D_l^0 \mathbf{V}_l \tag{IV.49}$$

$$\mathsf{Y}_{sl} = \nabla\phi_s D_l^1(\mathrm{e}^q - 1) \tag{IV.50}$$

Il faut néanmoins préciser que les termes non-linéaires traités de manière explicite $\delta\mathsf{X}$ et $\delta\mathsf{Y}$ ont une définition légèrement plus complexe :

$$\delta\mathsf{X}_l = \frac{1}{\delta_l R T_l}\left[\nabla(\phi_{\tilde{l}} - \phi_S)\cdot(\mathbf{V}_{\tilde{l}} - \mathbf{V}_l) + \nabla(\phi_{\tilde{l}-1} - \phi_S)\cdot(\mathbf{V}_l - \mathbf{V}_{\tilde{l}-1})\right] \tag{IV.51}$$

$$\delta\mathsf{Y}_l = \nabla(\phi_l - \phi_s) D_l^1(\mathrm{e}^q - 1) \tag{IV.52}$$

Maintenant que les opérateurs discrets sont définis, nous allons pouvoir appliquer une méthode de substitution pour résoudre le système implicite (IV.36)-(IV.40).

Résolution du problème implicite dans l'espace discret

Pour résoudre le système (IV.36)-(IV.40), la technique par substitution impose tout d'abord de calculer explicitement les membres de droite, notés dès lors $X^{\bullet\bullet,k}$. Dans l'espace discrétisé verticalement, et en utilisant les définitions de X_{sl} et Y_{sl}, le problème implicite prend la forme suivante :

$$\mathbf{V}_l^{k+1} + a_{jj}^y \Delta t \nabla\phi_s D_l^1(\mathrm{e}^{q^k} q^{k+1})_l = \mathbf{V}_l^{\bullet\bullet,k} \tag{IV.53}$$

$$w_l^{k+1} - g a_{jj} \Delta t D_{\tilde{l}}^1(\mathrm{e}^{q^k} q^{k+1})_l = w_l^{\bullet\bullet,k} \tag{IV.54}$$

$$T_l^{k+1} + \Delta t \frac{e^{q_l^k}}{C_v}\left[-a_{jj} D_l^0 w_{\tilde{l}}^{k+1} + a_{jj}^x \left(\frac{\nabla\phi_s}{g}\right)\cdot D_l^0 \mathbf{V}_l^{k+1}\right] = T_l^{\bullet\bullet,k} \tag{IV.55}$$

$$q_l^{k+1} + \Delta t \frac{C_p}{C_v}\frac{e^{q_l^k}}{H_l^k}\left[-a_{jj} D_l^0 w_{\tilde{l}}^{k+1} + a_{jj}^x \left(\frac{\nabla\phi_s}{g}\right)\cdot D_l^0 \mathbf{V}_l^{k+1}\right] = q_l^{\bullet\bullet,k} \tag{IV.56}$$

$$\pi_s^{k+1} = \pi_s^{\bullet\bullet,k} \tag{IV.57}$$

L'idée de la substitution consiste à remplacer la valeur implicite d'une variable dans une équation, par l'ensemble de l'équation de ladite variable. Il est donc possible de choisir en quelle variable

la résolution implicite peut être effectuée. Pour faciliter ce calcul, il est toujours préférable de choisir celle qui porte les conditions aux bords. Il faut remarquer que l'utilisation de la coordonnée masse permet d'énoncer simplement la condition élastique pour le toit du modèle, qui se traduit par une condition de Dirichlet (*ie :* $q_T^{k+1} = 0$). De même, la condition rigide à la surface peut s'exprimer facilement dans l'équation pronostique de q. Ainsi, la résolution du système implicite s'accomplit par la substitution de la variable $\mathbf{V}$ et w dans l'équation de q (IV.56). Cette manipulation fait apparaître l'opérateur laplacien vertical $\mathbf{L}_v$ défini précédemment, ainsi qu'un nouvel opérateur laplacien $\mathbf{L}_\phi \equiv D_l^0 \cdot D_l^1$ tel que :

$$\begin{aligned}
\mathbf{L}_{\phi l}(\mathrm{e}^q q)_l =& \frac{\pi_{\tilde{l}+1}\mathrm{e}^{q_{\tilde{l}+1}}}{\delta_l\delta\pi_{l+1}}(1-\epsilon_{l+1})(1-\epsilon_l)q_{l+2} \\
&+\left(\frac{1-\epsilon_l}{\delta_l\delta\pi_{l+1}}(\epsilon_{l+1}\pi_{\tilde{l}+1}\mathrm{e}^{q_{\tilde{l}+1}}-(1-\epsilon_l)\pi_{\tilde{l}}\mathrm{e}^{q_{\tilde{l}}})-\frac{\pi_{\tilde{l}}\mathrm{e}^{q_{\tilde{l}}}}{\delta_l\delta\pi_l}(1-\epsilon_l)(1-\epsilon_l-\epsilon_{l-1})\right)q_{l+1} \\
&-\left(\frac{\pi_{\tilde{l}}\mathrm{e}^{q_{\tilde{l}}}}{\delta_l\delta\pi_{l+1}}(1-\epsilon_l)\epsilon_l+\frac{\epsilon_l\pi_{\tilde{l}}\mathrm{e}^{q_{\tilde{l}}}-(1-\epsilon_{l-1})\pi_{\tilde{l}-1}\mathrm{e}^{q_{\tilde{l}-1}}}{\delta_l\delta\pi_l}(1-\epsilon_l-\epsilon_{l-1})\right. \\
&\qquad\left.-\frac{\pi_{\tilde{l}-1}\mathrm{e}^{q_{\tilde{l}-1}}}{\delta_l\delta\pi_{l-1}}(1-\epsilon_{l-1})\epsilon_{l-1}\right)q_l \\
&+\left(\frac{\pi_{\tilde{l}-1}\mathrm{e}^{q_{\tilde{l}-1}}}{\delta_l\delta\pi_l}\epsilon_{l-1}(1-\epsilon_l-\epsilon_{l-1})-\frac{\epsilon_{l-1}}{\delta_l\delta\pi_{l-1}}(\epsilon_{l-1}\pi_{\tilde{l}-1}\mathrm{e}^{q_{\tilde{l}-1}}-(1-\epsilon_{l-2})\pi_{\tilde{l}-2}\mathrm{e}^{q_{\tilde{l}-2}})\right)q_{l-1} \\
&+\frac{\pi_{\tilde{l}-2}\mathrm{e}^{q_{\tilde{l}-2}}}{\delta_l\delta\pi_{l-1}}\epsilon_{l-2}\epsilon_{l-1}q_{l-2}
\end{aligned} \tag{IV.58}$$

Les conditions aux bords imposent des définitions particulières aux frontières :

$$\begin{aligned}
\left[\mathbf{L}_{\phi l}(\mathrm{e}^q q)\right]_1 =& \frac{\pi_{\tilde{2}}\mathrm{e}^{q_{\tilde{2}}}}{\delta_1\delta\pi_2}(1-\epsilon_2)(1-\epsilon_1)q_3-\frac{\pi_{\tilde{1}}\mathrm{e}^{q_{\tilde{1}}}}{\delta_1}\epsilon_1(1-\epsilon_1)\left(\frac{1}{\delta\pi_1}+\frac{1}{\delta\pi_2}\right)q_1 \\
&+\frac{1}{\delta_1}\left(\frac{1-\epsilon_1}{\delta\pi_2}(\epsilon_2\pi_{\tilde{2}}\mathrm{e}^{q_{\tilde{2}}}-(1-\epsilon_1)\pi_{\tilde{1}}\mathrm{e}^{q_{\tilde{1}}})-\frac{\pi_{\tilde{1}}\mathrm{e}^{q_{\tilde{1}}}}{\delta\pi_1}(1-\epsilon_1)^2\right)q_2
\end{aligned} \tag{IV.59}$$

$$\begin{aligned}
\left[\mathbf{L}_{\phi l}(\mathrm{e}^q q)\right]_2 =& \frac{\pi_{\tilde{3}}\mathrm{e}^{q_{\tilde{3}}}}{\delta_2\delta\pi_3}(1-\epsilon_3)(1-\epsilon_2)q_4 \\
&+\left(\frac{1-\epsilon_2}{\delta_2\delta\pi_3}(\epsilon_3\pi_{\tilde{3}}\mathrm{e}^{q_{\tilde{3}}}-(1-\epsilon_2)\pi_{\tilde{2}}\mathrm{e}^{q_{\tilde{2}}})-\frac{\pi_{\tilde{2}}\mathrm{e}^{q_{\tilde{2}}}}{\delta_2\delta\pi_2}(1-\epsilon_2)(1-\epsilon_2-\epsilon_1)\right)q_3 \\
&-\left(\frac{\pi_{\tilde{2}}\mathrm{e}^{q_{\tilde{2}}}}{\delta_2\delta\pi_3}(1-\epsilon_2)\epsilon_2+\frac{1}{\delta_2\delta\pi_2}(\epsilon_2\pi_{\tilde{2}}-(1-\epsilon_1)\pi_{\tilde{1}\mathrm{e}^{q_{\tilde{1}}}})(1-\epsilon_2-\epsilon_1)\right. \\
&\left.-\frac{\pi_{\tilde{1}}\mathrm{e}^{q_{\tilde{1}}}}{\delta_2\delta\pi_1}(1-\epsilon_1)\epsilon_1\right)q_2 \\
&-\frac{\pi_{\tilde{1}}\mathrm{e}^{q_{\tilde{1}}}}{\delta_2}\left(\frac{\epsilon_1}{\delta\pi_1}-\frac{1-\epsilon_1-\epsilon_2}{\delta\pi_2}\right)\epsilon_1 q_1
\end{aligned} \tag{IV.60}$$

$$\begin{aligned}\left[\mathbf{L}_{\phi l}(\mathrm{e}^{q}q)\right]_{L-1} =& \left(\frac{1-\epsilon_{L-1}}{\delta_{L-1}\delta\pi_L}(\pi_{\tilde{L}}\mathrm{e}^{q_L}-(1-\epsilon_{L-1})\pi_{\tilde{L}-1}\mathrm{e}^{q_{\tilde{L}-1}})\right.\\ &\left.-\frac{\pi_{\tilde{L}-1}\mathrm{e}^{q_{\tilde{L}-1}}}{\delta_{L-1}\delta\pi_{L-1}}(1-\epsilon_{L-1})(1-\epsilon_{L-1}-\epsilon_{L-2})\right)q_L\\ &-\left(\frac{\pi_{\tilde{L}-1}\mathrm{e}^{q_{\tilde{L}-1}}}{\delta_{L-1}\delta\pi_L}(1-\epsilon_{L-1})\epsilon_{L-1}\right.\\ &+\frac{1}{\delta_{L-1}\delta\pi_{L-1}}(\epsilon_{L-1}\pi_{\tilde{L}-1}\mathrm{e}^{q_{\tilde{L}-1}}-(1-\epsilon_{L-2})\pi_{\tilde{L}-2}\mathrm{e}^{q_{\tilde{L}-2}})(1-\epsilon_{L-1}-\epsilon_{L-2})\\ &\left.+\frac{\pi_{\tilde{L}-2}\mathrm{e}^{q_{\tilde{L}-2}}}{\delta_{L-1}\delta\pi_{L-2}}(1-\epsilon_{L-2})\epsilon_{L-2}\right)q_{L-1}\\ &+\left(\frac{\pi_{\tilde{L}-2}\mathrm{e}^{q_{\tilde{L}-2}}}{\delta_{L-1}\delta\pi_{L-1}}\epsilon_{L-2}(1-\epsilon_{L-1}-\epsilon_{L-2})\right.\\ &\left.-\frac{\epsilon_{L-2}}{\delta_{L-1}\delta\pi_{L-2}}(\epsilon_{L-2}\pi_{\tilde{L}-2}\mathrm{e}^{q_{\tilde{L}-2}}-(1-\epsilon_{L-3})\pi_{\tilde{L}-3}\mathrm{e}^{q_{\tilde{L}-3}})\right)\mathrm{e}^{q_{\tilde{L}-2}}q_{L-2}\\ &+\frac{\pi_{\tilde{L}-3}\mathrm{e}^{q_{\tilde{L}-3}}}{\delta_{L-1}\delta\pi_{L-2}}\epsilon_{L-2}\epsilon_{L-3}q_{L-3} \end{aligned} \tag{IV.61}$$

$$\begin{aligned}\left[\mathbf{L}_{\phi l}(\mathrm{e}^{q}q)\right]_{L} =& \frac{\epsilon_{L-1}}{\delta_L}\left(\frac{1}{\delta\pi_L}(\pi_{\tilde{L}}\mathrm{e}^{q_L}-(1-\epsilon_{L-1})\pi_{\tilde{L}-1}\mathrm{e}^{q_{\tilde{L}-1}})-\frac{1}{\delta\pi_{L-1}}\pi_{\tilde{L}-1}\mathrm{e}^{q_{\tilde{L}-1}}(1-\epsilon_{L-1})\right)q_L\\ &-\frac{\epsilon_{L-1}}{\delta_L}\left(\frac{\pi_{\tilde{L}-1}\mathrm{e}^{q_{\tilde{L}-1}}}{\delta\pi_L}\epsilon_{L-1}+\frac{1}{\delta\pi_{L-1}}(\epsilon_{L-1}\pi_{\tilde{L}-1}\mathrm{e}^{q_{\tilde{L}-1}}-(1-\epsilon_{L-2})\pi_{\tilde{L}-2}\mathrm{e}^{q_{\tilde{L}-2}})\right)q_{L-1}\\ &+\frac{\pi_{\tilde{L}-2}\mathrm{e}^{q_{\tilde{L}-2}}}{\delta_L\delta\pi_{L-1}}\epsilon_{L-2}\epsilon_{L-1}q_{L-2} \end{aligned} \tag{IV.62}$$

L'utilisation de cet opérateur maintient l'ordre de précision du schéma de discrétisation spatiale du fait qu'il a été vérifié sur les modes propres du système que la qualité de la réponse de cet opérateur, ou de l'opérateur Laplacien naturel de la grille $\mathbf{L}_v$ est équivalente.

La définition du problème dans le corps du domaine apparaît relativement plus simple que pour la surface, pour laquelle les conditions aux bords interviennent. Ainsi, pour $l \in [1; L-1]$, l'équation d'Helmholtz à résoudre est la suivante :

$$q_l^{k+1} - \mathrm{e}^{q_l^k}\left(\frac{c_l^k\Delta t}{H_l^k}\right)^2\left[a_{jj}^2\mathbf{L}_{vl} + a_{jj}^x a_{jj}^y\left(\frac{\nabla\phi_s}{g}\right)^2\mathbf{L}_{\phi l}\right](\mathrm{e}^{q^k}q^{k+1})_l = q_l^{\bullet\bullet\bullet,k} \tag{IV.63}$$

avec $c_l^k = \sqrt{RT_l^k C_p/C_v}$ et :

$$q_l^{\bullet\bullet\bullet,k} = q_l^{\bullet\bullet,k} - \frac{C_p\mathrm{e}^{q_l^k}}{C_v H_l^k}\Delta t\left[a_{jj}D_l^0 w_{\tilde{l}}^{\bullet\bullet,k} - a_{jj}^x\frac{\nabla\phi_s}{g}\cdot D_l^1\mathbf{V}_l^{\bullet\bullet,k}\right]$$

Traitement particulier à la surface

La condition rigide à la surface $gw_s = \mathbf{V}_s \cdot \nabla\phi_s$ s'insère directement durant la résolution dans cette équation implicite en écrivant l'équation (IV.56) sur le dernier niveau L :

$$q_L^{k+1} + e^{q_L^k}\Delta t \frac{C_p}{C_v}\left(-\frac{a_{jj}}{H_L^k}\frac{w_{\tilde{L}}^{k+1} - w_{\tilde{L}-1}^{k+1}}{\delta_L} + a_{jj}^x \frac{\nabla\phi_s}{RT_l}\cdot D_L^0 \mathbf{V}_l^{k+1}\right) = q_L^{\bullet\bullet,k}$$

De là, il ne reste plus qu'à utiliser la définition des opérateurs pour ainsi avoir la dernière ligne de l'équation matricielle :

$$q_L^{k+1} - \mathrm{e}^{q_L^k}\left(\frac{c_L^k \Delta t}{H_L^k}\right)^2 \left[a_{jj}^2 \mathbf{L}_{vL} + a_{jj}^x a_{jj}^y \left(\frac{\nabla\phi_s}{g}\right)^2 \mathbf{L}_{\phi L} + \frac{a_{jj}a_{jj}^y}{\delta_L}\left(\frac{\nabla\phi_s}{g}\right)^2 D_L^1\right](\mathrm{e}^{q^k} q^{k+1})_L = q_L^{\bullet\bullet\bullet,k} \tag{IV.64}$$

avec enfin :

$$q_L^{\bullet\bullet\bullet,k} = q_L^{\bullet\bullet,k} - \mathrm{e}^{q_L^k}\frac{\Delta t}{H_L^k}\frac{C_p}{C_v}\left[a_{jj} D_l^0 w_{\tilde{L}}^{\bullet\bullet,k} - a_{jj}^x\left(\frac{\nabla\phi_s}{g}\right)\cdot D_l^1 \mathbf{V}_L^{\bullet\bullet,k} - \frac{a_{jj}}{\delta_L}\left(\frac{\nabla\phi_s}{g}\right)\cdot \mathbf{V}_L^{\bullet\bullet,k}\right]$$

Afin de bien mettre en évidence la recherche du point fixe réalisée par cette méthode quasi-Newton, nous définissons de manière symbolique l'opérateur du problème de Helmholtz discret (IV.63)-(IV.64) ci-dessous par $\mathcal{H}_k$ tel que :

$$\mathcal{H}_k(q^{k+1}) = q^{\bullet\bullet\bullet,k} \tag{IV.65}$$

La méthode de résolution de (IV.65) est fonction de la valeur des coefficients a_{jj}^x et a_{jj}^y. Dans le cas où l'un de ces deux coefficients est nul, alors l'opérateur discret $\mathcal{H}^k$ à inverser est une matrice tridiagonale, le problème se résout donc par un algorithme de double descente classique (voir Appendix C). Par commodité, cette version est nommée VITE (Verticalement Implicite Tridiagonale Équation). Dans ce cas, l'orographie est traitée soit de manière explicite, soit, selon la méthode Mixed, c'est-à-dire que les termes X_s et Y_s sont respectivement traités par une substitution implicite comme la divergence horizontale du vent ou le gradient de pression horizontal. En terme de traitement par tableaux de Butcher, le traitement Mixed de ces termes orographiques signifie les équivalences suivantes : $\{c^x, \mathcal{A}^x, b^x\} \equiv \{c^u, \mathcal{A}^u, b^u\}$ et $\{c^y, \mathcal{A}^y, b^y\} \equiv \{c^p, \mathcal{A}^p, b^p\}$. En conséquence, du fait de ces substitutions implicites, le problème à résoudre demeure tridiagonal.

Dans le cas où le traitement des termes orographiques est implicite ; autrement dit, si les termes X_s et Y_s sont traités dans le même tableau implicite que les termes d'ajustement verticaux, (*ie :* $\{c^x, \mathcal{A}^x, b^x\} \equiv \{c^y, \mathcal{A}^y, b^y\} \equiv \{c, \mathcal{A}, b\}$), l'opérateur discret discret $\mathcal{H}^k$ à inverser est une matrice penta-diagonale (dont les coefficients diagonaux sont non-nuls). La méthode d'inversion employée

dans ce cas reste directe, et peut être considérée comme une généralisation de l'algorithme de double descente (*cf* Appendix C pour plus de détails). Ce traitement est appelé par l'abréviation VIPE (Verticalement Implcite Penta-diagonale Équation).

La procédure permettant d'obtenir l'ensemble de l'état $X^{(k+1)}$ peut être résumée de la manière suivante. La pression de surface π_s^{k+1} est la première variable à être déterminée explicitement via (IV.40). Une fois que la résolution de (IV.63) effectuée, la solution q^{k+1} est injectée dans les équations du mouvement (IV.36) et (IV.37) pour obtenir $\mathbf{V}^{(k+1)}$ et $w^{(k+1)}$. Enfin, il ne reste plus qu'à déterminer T^{k+1} par substitution dans (IV.38).

Comme nous effectuons une résolution itérative de problème inverse (IV.34), nous ne faisons qu'approcher la solution exacte de cette équation par une suite d'états. Il est donc nécessaire de déterminer un seuil de tolérance à partir duquel la solution calculée est considérée comme suffisamment proche de la solution exacte pour arrêter les calculs.

Critères de convergence

Pour assurer que l'état calculé converge vers la solution, il est nécessaire de vérifier deux critères de convergence. Le premier impose que la suite q^k approche au moins une solution du problème d'Helmholtz (IV.65) à la tolérance fixée $\epsilon_1 = 10^{-10}$. Pour cela, on définit le résidu de la k-ième itération R^k comme :

$$R^k = \mathcal{H}_k(q^k) - q^{\bullet\bullet\bullet,k}$$

On considère que la solution obtenue au bout de la k-ième est suffisamment proche de la solution recherchée si la norme Euclidienne [2] du k-ième résidu satisfait :

$$\left|\left|R^k\right|\right|_2 \leq \epsilon_1 \left|\left|R^0\right|\right|_2, \tag{IV.66}$$

où de manière transparente $||R^0||_2$ est la norme Euclidienne du résidu de l'itération zéro calculée au début du processus itératif. Le second critère à vérifier, est l'unicité de la solution. Pour ce faire, la suite des q^k doit converger. Comme c'est un vecteur de $\mathbb{R}^L$, et que $\mathbb{R}$ est un espace complet (par définition), alors, il suffit que q^k soit une suite de Cauchy [3]. Dans le mesure où la valeur de q^k est, de manière générale, très faible (de l'ordre de 10^{-3}), nous avons choisi d'opérer le critère de convergence sur la suite des e^{q^k}, de sorte que :

$$\left|\left|\mathrm{e}^{q^{k+1}} - \mathrm{e}^{q^k}\right|\right|_2 \leq \epsilon_2 \left|\left|\mathrm{e}^{q^1} - \mathrm{e}^{q^0}\right|\right|_2, \tag{IV.67}$$

avec la tolérance ϵ_2 prise égale à 10^{-7} (valeur déterminée empiriquement). Enfin, une dernière sécurité consiste à vérifier que le nombre maximum d'itérations autorisées soit toujours inférieure à $N_{iter} = 10$. Si les deux critères de convergence (IV.66) et (IV.67) sont respectés ou que k égale à N_{iter}, alors la recherche du point fixe (IV.65) s'arrête. Dans le cas où le schéma est stable, de tels critères d'arrêt assurent donc la bonne résolution du quasi-Newton, et donc une bonne inversion du problème non-linéaire initiale.

2. Euclide (environ 300 av. J-C) : mathématicien de la Grèce antique
3. Augustin Louis, baron Cauchy (1789-1857) : mathématicien français

Maintenant que deux approches VITE et VIPE ont été décrites et que leur faisabilité a été démontrée dans le cas d'un modèle non-linéaire, il reste à prouver que l'approche VIPE présente un véritable avantage en terme de stabilité par rapport à l'approche usuelle VITE. Pour cela, la section suivante se tourne, là encore, vers l'analyse du système linéaire mais, cette fois-ci, dans l'espace discrétisé verticalement.

3 Étude de stabilité discrète sur la verticale

Pour modéliser l'impact de l'orographie sur la stabilité des schémas retenus d'après le chapitre précédent, nous allons réaliser une analyse sur un système linéaire ayant une orographie. Cette analyse théorique est, par essence, très éloignée d'un cadre physique réaliste. Elle vise uniquement à donner une indication, une tendance générale, qui doit pouvoir être confirmée par des expériences numériques. Nous présentons donc la différence entre les comportements de la stabilité des versions VITE et VIPE, tout en restant critique, et en sachant pertinemment que ces résultats sont très optimistes.

Définition du système linéaire continu

L'état de référence à partir duquel doit s'opérer la linéarisation possède globalement les mêmes hypothèses que $\overline{X}$ (pour rappel, il est au repos, hydrostatique, isotherme ($\overline{T}$), uniforme, stationnaire en 2D plan vertical et en coordonnée σ). Pour mesurer l'impact de l'orographie sur la stabilité des schémas, un des moyens est de supposer que le forçage de cette orographie est constant. Cette démarche définie par Bénard (2005) [8], nécessite donc que l'état de référence possède une orographie telle que la hauteur de la surface z_s soit :

$$z_s(x) = sx$$

où s est le pourcentage de la pente compris dans l'intervalle $[0, 1]$.

L'introduction de cette orographie met en exergue une condition sur la pression hydrostatique $\overline{\pi}_s$ afin de maintenir l'état de référence au repos. En effet, si nous imposons toutes ces hypothèses, l'équation du mouvement (8) devient :

$$\begin{aligned} & R\overline{T}\frac{\nabla\overline{\pi}_s}{\overline{\pi}_s} + \nabla\phi_s = 0 \\ \iff \quad & \nabla \ln \overline{\pi}_s = \frac{s}{\overline{H}} \end{aligned} \tag{IV.68}$$

Grâce à cette relation, nous pouvons effectuer le reste de la linéarisation dans un modèle plan

vertical $(x-\sigma)$:

$$\partial_t u + R\left(\mathcal{G}\nabla T - \frac{s}{\overline{\overline{H}}}T\right) + R\overline{T}\left(\mathrm{I}-\mathcal{G}\right)\nabla q + gs\left(\tilde{\partial}+\mathrm{I}\right)q + \frac{R\overline{T}}{\overline{\pi}_s}\nabla\pi_s = 0 \tag{IV.69}$$

$$\partial_t w - g(\tilde{\partial}+\mathcal{I})q = 0 \tag{IV.70}$$

$$\partial_t T + \frac{R\overline{T}}{C_v}\left(\nabla u - \frac{1}{\overline{\overline{H}}}\tilde{\partial}w + \frac{s}{\overline{\overline{H}}}\tilde{\partial}u\right) = 0 \tag{IV.71}$$

$$\partial_t q + \frac{C_p}{C_v}\left(\nabla u - \frac{1}{\overline{\overline{H}}}\tilde{\partial}w + \frac{s}{\overline{\overline{H}}}\tilde{\partial}u\right) - \mathcal{S}\nabla u - \frac{s}{\overline{\overline{H}}}\left(\mathrm{I}-\mathcal{S}\right)u = 0 \tag{IV.72}$$

$$\partial_t \pi_s + \overline{\pi}_s\mathcal{N}\left(\nabla - \frac{s}{\overline{\overline{H}}}\right)u = 0 \tag{IV.73}$$

Écrit ainsi, il apparaît que les termes résiduels $\delta\mathsf{X}$ et $\delta\mathsf{Y}$ sont nuls. En revanche, nous pouvons identifier la valeur de ces termes croisés à la surface :

$$\mathsf{X}_s = \frac{s}{\overline{\overline{H}}}\tilde{\partial}u$$
$$\mathsf{Y}_s = gs\left(\tilde{\partial}+\mathcal{I}\right)q$$

Maintenant que le système est bien défini, il faut savoir comment l'ensemble des termes supplémentaires sont à traiter lors de l'utilisation des schémas RK-IMEX.

Mise en place des matrices d'amplifications dans l'espace discrétisé verticalement

Contrairement aux analyses déjà réalisées dans les chapitres précédents, la présence d'orographie (et donc de la condition au bord inférieur) nous empêche de pouvoir réaliser l'analyse sur le système non-borné.. Nous utilisons donc la méthode de Von Neumann pour effectuer l'étude de stabilité en considérant que les solutions sont de la forme :

$$\Psi_l(x,t) = \hat{\Psi}_l(t)\exp[\hat{\imath}kx], \quad \text{pour} \quad l \in [\![1;L]\!]$$

avec k le nombre d'ondes horizontales du mode de Fourier étudié. Dans la mesure où les grands nombres d'ondes (moins bien résolus dans un modèle) sont ceux qui sont les plus problématiques pour la stabilité des schémas, nous nous plaçons directement dans le cas de la plus petite onde résolue par le modèle défini par $k = k_{\max} = \pi/\Delta x$.

Dans la direction verticale, nous utilisons les opérateurs discrétisés verticalement (définis ci-avant), qui sont semblables soit à des matrices carrées ($L\times L$ ou $(L+1)\times(L+1)$, avec L le nombre de niveaux) soit à des matrices rectangulaires ($L\times(L+1)$ ou $(L+1)\times L$). Les opérateurs sont définis avec la coordonnée σ définie de sorte que l'épaisseur de la maille soit de Δz. On définit alors le rapport d'aspect des résolutions $r = \Delta x/\Delta z$. En effet, l'étude sur le système des ondes acoustiques 2D montre que la stabilité dépend, grossièrement, du produit sr. Ainsi, le but de ces analyses est de montrer l'impact sur la stabilité , en fonction des paramètres s et r, du nombre de Courant et du traitement VITE ou VIPE.

Afin d'accroître la stabilité en présence d'orographie, l'idée est ici de traiter les termes X_s et Y_s de manière implicite. Pour savoir comment traiter les autres termes qui apparaissent du fait

de la présence de l'orographie, il suffit de savoir de quels termes linéarisés ils proviennent. En effet, d'après le deuxième chapitre, il a été démontré que pour avoir une équation de structure discrète analogue à celle du système continu, il est nécessaire de traiter certains termes au même instant temporel. Par exemple, le terme d'auto-convection $\dot{\pi}/\pi$ doit être traité comme celui de la divergence horizontale du vent. De ce fait le terme d'ordre zéro lié à l'orographie dans l'équation (IV.72) doit être traité par le schéma associé à l'opérateur $\mathscr{U}$. Pour des raisons similaires, le terme d'ordre zéro relié à la pente s de l'équation (IV.69), issu de la linéarisation des termes du gradient de pression, doit être traité par le même tableau de Butcher que celui associé à l'opérateur $\mathscr{P}$.

Les opérateur discrets $\mathbf{I}'$, $\mathbf{U}$, $\mathbf{P}$, $\mathbf{X}$ et $\mathbf{Y}$ sont définis par :

$$\mathbf{I}' = \begin{pmatrix} \mathbf{0} & \hat{\mathbf{0}} & \mathbf{0} & \mathbf{0} & 0 \\ \breve{\mathbf{0}} & \tilde{\mathbf{0}} & \breve{\mathbf{0}} & gD_{\bar{l}}^{1} & \breve{0} \\ \frac{R\overline{T}}{C_v}\frac{s}{\delta_L \overline{H}}\mathbf{C}_b & \frac{R\overline{T}}{C_v \overline{H}}D_l^0 & \mathbf{0} & \mathbf{0} & 0 \\ \frac{C_p}{C_v}\frac{s}{\delta_L \overline{H}}\mathbf{C}_b & \frac{C_p}{C_v \overline{H}}D_l^0 & \mathbf{0} & \mathbf{0} & 0 \\ 0 & \hat{0} & 0 & 0 & 0 \end{pmatrix}$$

$$\mathbf{U} = -\begin{pmatrix} \mathbf{0} & \hat{\mathbf{0}} & \mathbf{0} & \mathbf{0} & 0 \\ \breve{\mathbf{0}} & \tilde{\mathbf{0}} & \breve{\mathbf{0}} & \breve{\mathbf{0}} & \breve{0} \\ \frac{R\overline{T}}{C_v}\hat{\imath}k\mathrm{I} & \hat{\mathbf{0}} & \mathbf{0} & \mathbf{0} & 0 \\ \frac{C_p}{C_v}\hat{\imath}k\mathrm{I} - \mathbf{S}\hat{\imath}k - \frac{s}{\overline{\overline{H}}}(\mathrm{I} - \mathbf{S}) & \hat{\mathbf{0}} & \mathbf{0} & \mathbf{0} & 0 \\ \overline{\pi}_s \mathbf{N}\left(\nabla - \frac{s}{\overline{\overline{H}}}\right) & 0 & 0 & 0 & 0 \end{pmatrix}$$

$$\mathbf{P} = -\begin{pmatrix} \mathbf{0} & \hat{\mathbf{0}} & R\left(\mathbf{G}\hat{\imath}k - \frac{s}{\overline{\overline{H}}}\mathrm{I}\right) & R\overline{T}(\mathrm{I} - \overline{\mathbf{G}})\hat{\imath}k\mathrm{I} & \frac{R\overline{T}}{\overline{\pi}_s}\hat{\imath}k\mathrm{I} \\ \breve{\mathbf{0}} & \tilde{\mathbf{0}} & \breve{\mathbf{0}} & \breve{\mathbf{0}} & \breve{0} \\ \mathbf{0} & \hat{\mathbf{0}} & \mathbf{0} & \mathbf{0} & 0 \\ \mathbf{0} & \hat{\mathbf{0}} & \mathbf{0} & \mathbf{0} & 0 \\ 0 & \hat{0} & 0 & 0 & 0 \end{pmatrix}$$

$$\mathbf{X} = -\begin{pmatrix} \mathbf{0} & \hat{\mathbf{0}} & \mathbf{0} & \mathbf{0} & 0 \\ \breve{\mathbf{0}} & \tilde{\mathbf{0}} & \breve{\mathbf{0}} & \breve{\mathbf{0}} & \breve{0} \\ \frac{R\overline{T}}{C_v}\frac{s}{\overline{\overline{H}}}D_l^0 & \hat{\mathbf{0}} & \mathbf{0} & \mathbf{0} & 0 \\ \frac{C_p}{C_v}\frac{s}{\overline{\overline{H}}}D_l^0 & \hat{\mathbf{0}} & \mathbf{0} & \mathbf{0} & 0 \\ 0 & \hat{0} & 0 & 0 & 0 \end{pmatrix} \quad \mathbf{Y} = -\begin{pmatrix} \mathbf{0} & \hat{\mathbf{0}} & \mathbf{0} & gsD_l^1 & 0 \\ \breve{\mathbf{0}} & \tilde{\mathbf{0}} & \breve{\mathbf{0}} & \breve{\mathbf{0}} & \breve{0} \\ \mathbf{0} & \hat{\mathbf{0}} & \mathbf{0} & \mathbf{0} & 0 \\ \mathbf{0} & \hat{\mathbf{0}} & \mathbf{0} & \mathbf{0} & 0 \\ 0 & \hat{0} & 0 & 0 & 0 \end{pmatrix}$$

avec I la matrice identité de taille $L \times L$, et les matrices $\mathbf{0}$, $\tilde{0}$, $\breve{0}$ et $\hat{0}$ qui sont les matrices nulles respectivement de dimension $L \times L$, $(L+1) \times (L+1)$, $L \times (L+1)$ et $(L+1) \times L$. Les mêmes notations sont utilisées à la fois pour la dernière colonne de ces matrices, et pour leur dernière ligne. De plus, la matrice $\mathbf{C}_b$ de dimension $L \times L$ contient l'information de la condition à la surface du problème. Elle est donc nulle partout, sauf au point (L, L), avec $\mathbf{C}_{bLL} = 1$. Enfin, comme l'advection est absente de cette étude, l'opérateur $\mathbf{E}$ subissant un traitement purement explicite est nul.

Finalement, les matrices d'amplifications du schéma RK-IMEX HEVI sont données par les relations de récurrences matricielles :

$$\begin{aligned} A^{(j)} =& \mathbf{1} + \Delta t \sum_{i=1}^{j} a_{ji}^{x} \mathbf{X} \cdot A^{(i)} + \Delta t \sum_{i=1}^{j} a_{ji}^{y} \mathbf{Y} \cdot A^{(i)} + \sum_{i=1}^{j} a_{ji} \mathbf{I}' \cdot A^{(i)} \\ &+ \sum_{i=1}^{j} a_{ji}^{u} \mathbf{U} \cdot A^{(i)} + \sum_{i=1}^{j} a_{ji}^{p} \mathbf{P} \cdot A^{(i)} \end{aligned} \tag{IV.74}$$

$$\begin{aligned} A =& \mathbf{1} + \Delta t \sum_{i=1}^{\nu} b_{j}^{x} \mathbf{X} \cdot A^{(j)} + \Delta t \sum_{i=1}^{\nu} b_{j}^{y} \mathbf{Y} \cdot A^{(j)} + \sum_{i=1}^{\nu} b_{j} \mathbf{Y} \cdot A^{(j)} \\ &+ \sum_{i=1}^{\nu} b_{j}^{u} \mathbf{U} \cdot A^{(j)} + \sum_{i=1}^{\nu} b_{j}^{p} \mathbf{P} \cdot A^{(j)} \end{aligned} \tag{IV.75}$$

où $\mathbf{1}$ désigne l'opérateur discret identité dans l'espace du vecteur d'état discret.

Là encore, le schéma est stable si, et seulement si, l'ensemble des valeurs propres de A ont un module inférieur ou égal à 1. Ceci est équivalent au fait que le coefficient d'amplification Γ (la norme maximale des valeurs propres) doit être inférieur à 1.

Résultat de l'étude de stabilité en présence d'une pente

La Figure IV.3 montre le coefficient d'amplification de la version UFpreF en fonction de s et C_* pour plusieurs valeurs de r. D'après le chapitre précédent, une condition nécessaire à la stabilité du schéma est que $C_* \leq 2$. De manière générale, la présence d'orographie déstabilise fortement les schémas numériques, et c'est pour cela que les petites instabilités (*ie :* $\Gamma \in]1; 1{,}05]$), qui sont susceptibles de disparaître avec l'utilisation d'une légère diffusion, sont marquées d'une isoligne blanche. Dans le cas VITE, il apparaît clairement qu'il existe une relation hyperbolique entre le nombre de Courant maximum assurant la stabilité, et le produit sr. Cette dépendance est très problématique car les modèles de PNT sont définis avec rapport r dépassant largement 100 (*ex :* AROME $r = \Delta x / \Delta z_{\min} \approx 250$). Cette analyse révèle l'importance de traiter ces termes de manière particulière afin de pouvoir être employés en opérationnel. Pour cette version du Trap2(2,3,2)(-1), le gain de stabilité est très important grâce à une résolution VIPE. Dans le cas d'un rapport $r = 1$ (figure b)), il semble que la contrainte liée à l'orographie disparaisse totalement. Dans le cas où $r = 10$ (figure c), les petites instabilités semblent facilement effaçables avec l'utilisation d'une diffusion. De même, les coefficients d'amplification pour $r = 100$ sont relativement bien moins inférieurs, dans le cas VIPE. Ceci laisse penser que l'utilisation de termes diffusifs seraient largement plus efficaces pour stabiliser la version VIPE que celle de VITE.

Les différences de comportements de la stabilité du schéma Trap2-Mixed, illustré à la Figure IV.4 sont bien plus ténues que dans le cas UFpreF. Comme dans le cas de l'advection, il semble que l'équilibre, créé par la distribution de termes implicites dans les équations du mouvement et de la pression, permette de maintenir la stabilité par rapport au cas sans orographie. Ce schéma confirme néanmoins que le traitement VIPE est plus stable (notamment pour $r = 10$). Dans le cas où le schéma est instable les figures (k et l) montrent que le coefficient d'amplification reste plus faible avec un traitement VIPE, ce qui permet l'utilisation de diffusion plus faible.

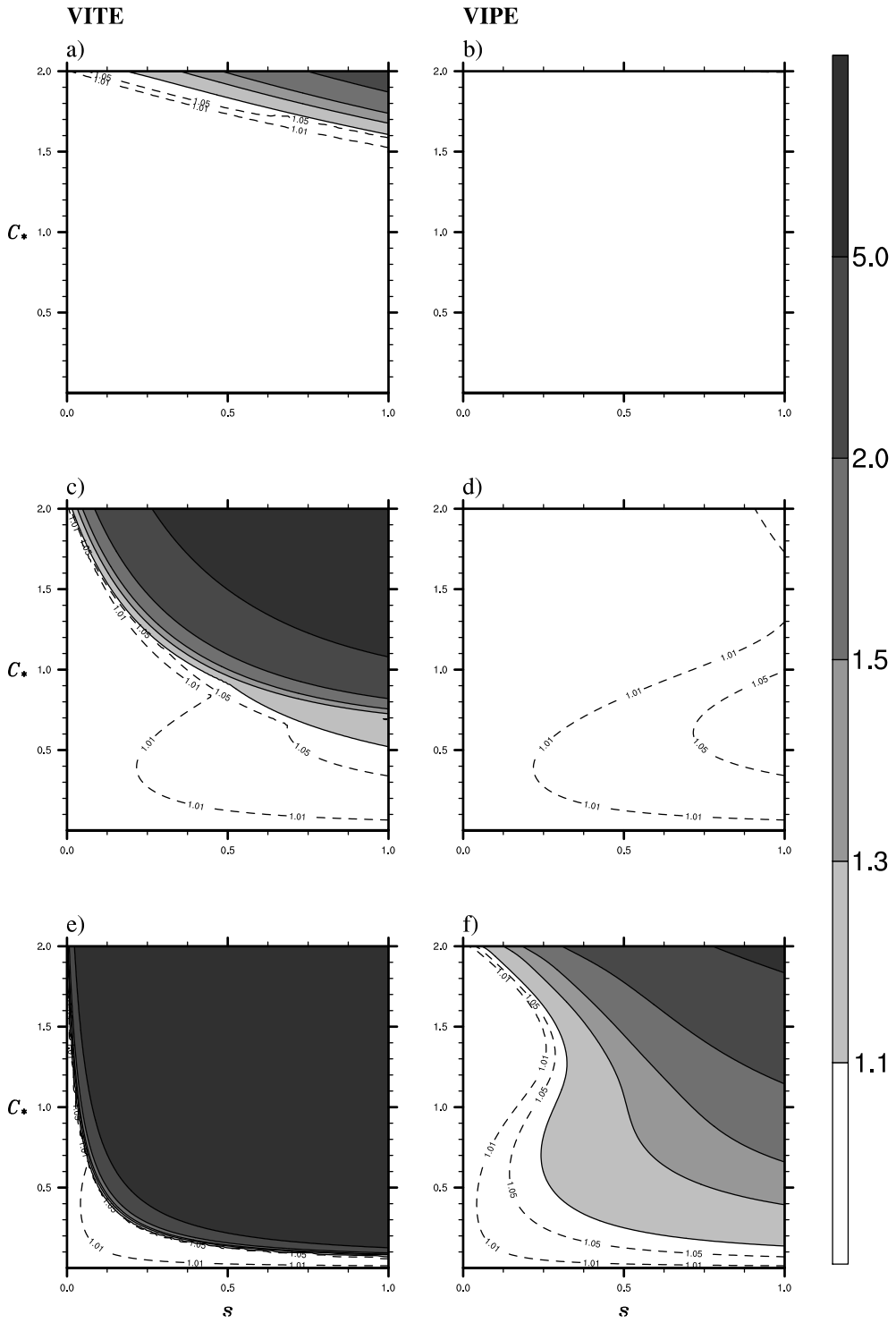


FIGURE IV.3 – *Coefficient d'ampliation du schéma Trap2(2,3,2)(-1) UFpreF en fonction de s et C_{a_x} pour $r = 1$ en haut, $r = 10$ au milieu et $r = 100$.*

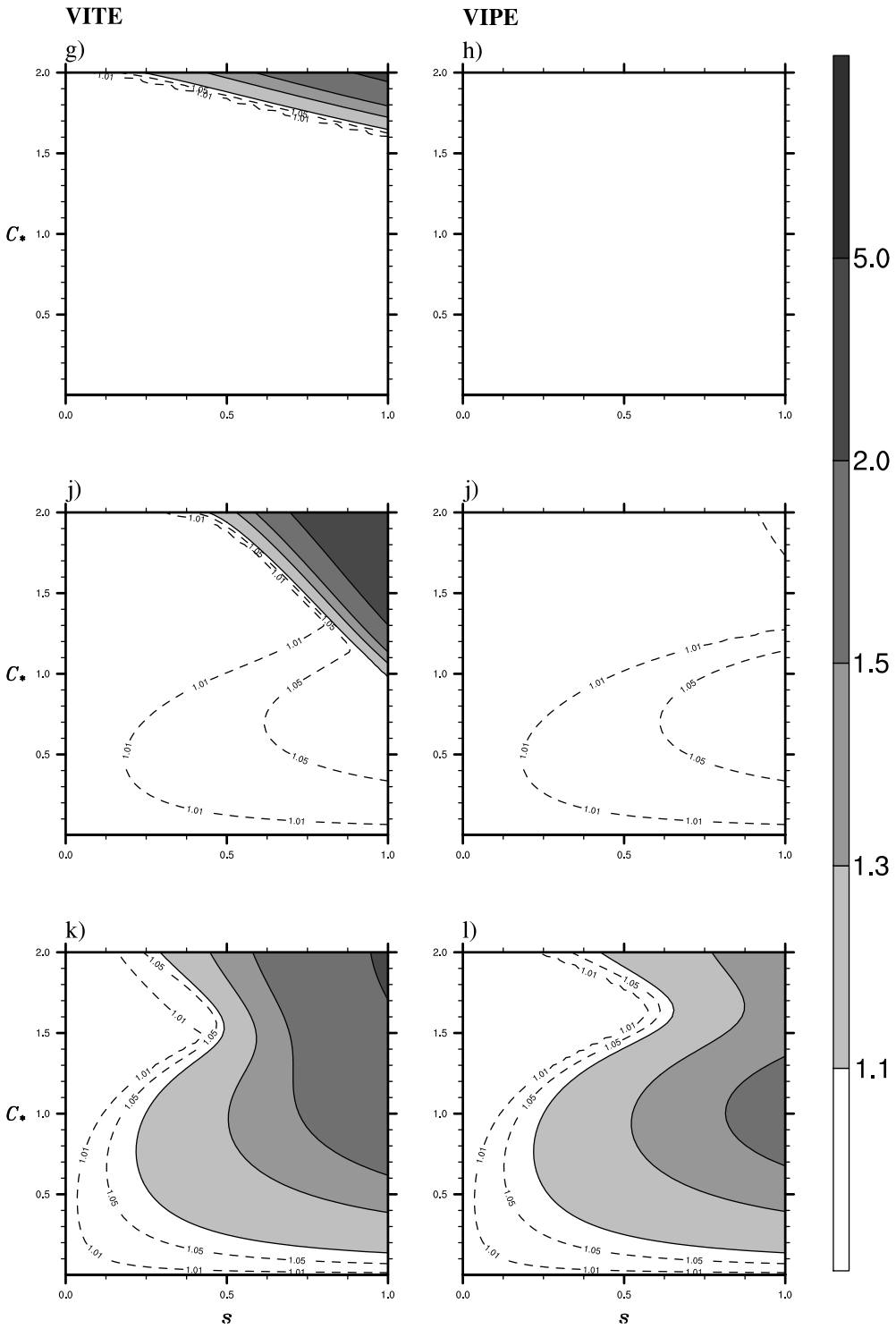


FIGURE IV.4 – *Même graphique que IV.3, mais pour le schéma Trap2-Mixed*

Par ailleurs, la comparaison entre le traitement Mixed-VITE et le traitement VIPE avec un schéma UFpreF offre un nouvel argument en faveur des schémas Mixed. En effet, alors que l'utilisation de méthodes d'inversion de matrices penta-diagonales sont plus onéreuses qu'une simple inversion tri-diagonale, il semble que l'équilibre entre la stabilité et l'économie de calculs plaide, une nouvelle fois, en faveur des méthodes aux quatre tableaux de Butcher. Par rapport au deux cas VIPE, il semble que les schémas Mixed soient, là encore, légèrement plus stables notamment en comparant pour un rapport de $r = 100$ (figures f et l). Mais ce comportement ne semble pas toujours vérifié. En effet, pour le cas $r = 10$, (figures d et j), et pour des pentes $s = 1$ et pour des nombres de Courant $C_* \in [1; 1{,}5]$, il semble que les schémas Mixed soient moins stables. Ces cas particuliers restent suffisamment dérisoires pour considérer que si le choix d'un traitement VIPE est effectué, il semble plus efficace de l'appliquer au schéma Trap2-Mixed.

Ces premières analyses confirment les intuitions à la fois sur le caractère instable des termes orographiques, mais aussi sur le gain de stabilité à les traiter de manière implicite. Ces résultats restent vrais quelque que soit le schéma utilisé. Mais à ce stade, ces résultats ne sont pas directement transposables pour un modèle non-linéaire, et c'est pourquoi, des tests avec un modèle numérique doivent être réalisés afin de les confirmer.

4 Discussion

Les études de stabilité menées dans ce chapitre, pour le cadre particulier d'un forçage orographique constant, ont permis de montrer plusieurs choses. La première confirme que, de manière générale, les termes orographiques imposent de nouvelles conditions sur la stabilité. Un traitement implicite de ces termes verticaux semble toujours relaxer cette nouvelle contrainte. Ce nouveau traitement des termes orographiques peut être appliqué pour n'importe quel schéma HEVI faisant apparaître ces termes, y compris les méthodes au pas de temps fractionné HEVI. Il faut noter que ces termes sont également présents pour d'autres systèmes de coordonnées, notamment la coordonnée hauteur de type Gal-Chen & Somerville (1975) [23]. Ainsi, cette idée de traitement implicite peut s'appliquer à de nombreux modèles de la PNT pour accroître l'efficacité des algorithmes en présence de pentes.

Dans le prochain chapitre, nous allons réaliser quelques expériences numériques pour confirmer une bonne partie des analyses réalisées au cours de ces différents chapitres d'analyses théoriques.

Chapitre V

Validations expérimentales

Dans les précédents chapitres, il a toujours était question d'analyse de stabilité pour déterminer le schéma autorisant le plus grand pas de temps possible. Il a été établi que les méthodes au pas de temps fractionné pouvaient souffrir d'un problème de stabilité pour de forts jets d'altitude ce qui nous a conduit à étudier les schémas RK-IMEX. De là, nous avons mis au point un schéma Trap2-Mixed qui semble avoir les qualités requises en termes de stabilité et de précision pour une utilisation en PNT. Enfin, dans le chapitre précédent, nous avons suggéré de traiter de manière implicite les termes orographiques, issus d'une utilisation d'une coordonnée verticale épousant le terrain, ce qui pourrait accroître la stabilité des schémas HEVI.

Dans la mesure où les analyses ont été réalisées dans un cadre théorique très éloigné de cas réalistes, il faut confronter ces résultats à des expériences numériques avec un modèle laboratoire. Grâce à cette approche expérimentale, nous chercherons à répondre aux questions suivantes : dans quelle mesure les études précédentes sur la stabilité des schémas sont-elles vérifiées par l'expérience numérique ? Les nouveaux schémas élaborés ont-ils réellement la précision attendue ? Le schéma Trap2-Mixte est-il toujours le meilleur candidat ? Le traitement implicite des termes orographiques (VIPE) apporte-t-il véritablement un gain de stabilité par rapport au traitement explicite (VITE) habituellement employé ?

Le but de ce chapitre final est de vérifier les comportements décrits par les études précédentes. Pour cela, nous avons simulé l'écoulement de différents fluides par différents modèles : la méthode Trap2(2,3,2)(-1) (avec ses différentes variantes UFpreF et Mixed et pour le traitement des termes orographiques), le schéma K(M)-Split, la méthode SI actuellement utilisée dans AROME et une méthode explicite Runge-Kutta-4 classique (RK4) qui nous servira de référence pour comparer qualitativement les résultats de l'intégration des différents schémas. Les discrétisations spatiales sont les mêmes pour tous ces modèles, et sont décrites par le chapitre précédent.

1 Définition de l'état de base

Les cas tests présentés dans ce chapitre sont issus de plusieurs expériences standards dans des cadres idéalisés bi-dimensionnels très souvent proposés dans la littérature scientifique pour valider des schémas de discrétisations spatio-temporelles. On y trouve le cas de la propagation horizontale des ondes de gravité dans un régime d'écoulement non-hydrostatique, le cas des écoulements contrôlés par la flottabilité atmosphérique, et bien d'autres tests d'écoulements orographiques pour diverses valeurs du nombre de Froude [1]. Tous les tests présentés ici sont effectués en absence de la force de Coriolis. Le but est de démontrer la hiérarchie par rapport à la stabilité des schémas que nous avons élaborés, et leur capacité à reproduire correctement les caractéristiques des différentes solutions de ces différents cas tests.

Les états initiaux de chacune des expériences ci-après sont définis comme la somme d'un état de base et d'une perturbation. Ces perturbations seront décrites avant chaque expérience. Dans ce paragraphe, nous définissons à la fois l'état de base et la géométrie générale des expériences.

L'état de base, noté $\overline{X}$, est en équilibre hydrostatique $\overline{q} = 0$ (et donc sans déplacement vertical $\overline{w} = \overline{\dot{\eta}} = 0$), avec un écoulement uniforme $\overline{U}$ le long de la direction-x et sans orographie. La température est définie par rapport à une fréquence de Brunt-Väisälä $\overline{N}$ constante, de sorte que le profil thermique de l'atmosphère de base est donné par :

$$\overline{T}(z) = \begin{cases} T_s - \dfrac{g}{C_p}(z - z_s), & \text{si} \quad \overline{N} = 0 \\ T_s \left\{ \left(1 - \dfrac{N_s^2}{\overline{N}^2}\right) \exp\left[\dfrac{\overline{N}^2}{g}(z - z_s)\right] + \dfrac{N_s^2}{\overline{N}^2} \right\}, & \text{sinon} \end{cases} \tag{V.1}$$

avec $N_s^2 = g^2/(C_p T_s)$, où T_s désigne la température de surface et z_s est la hauteur de l'orographie. En définissant la température potentielle à la surface $\theta_s = T_s(\overline{p}_s/p_{00})^{R/C_p}$. Comme la pression de surface de l'état de base est définie par $\overline{p}_s = p_{00} = 1000\,\text{hPa}$, alors, $T_s = \theta_s$. En suivant les définitions précédentes, la température potentielle se calcule ainsi :

$$\overline{\theta}(z) = \theta_s \exp[\overline{N}^2(z - z_s)/g]$$

Pour le besoin des expériences, la vitesse du son maximale est donnée par $\overline{c}_s = \sqrt{(C_p/C_v)R|\overline{T}|_\infty}$.

Sur la verticale, la grille choisie est celle de Lorenz où la vitesse verticale covariante $\dot{\eta}$ et contravariante w, ainsi que la pression hydrostatique π et le géopotentiel $\phi = gz$ sont définis sur les interfaces $\tilde{l}$, alors que les autres variables sont définies sur les niveaux $l \in [0; L]$. L est l'indice du niveau le plus bas, et $\tilde{L}$ est l'indice du dernier niveau correspondant à la surface (voir Figure IV.2). La coordonnée verticale est une coordonnée masse sigma défini par $\sigma_{\tilde{l}} = \pi_{\tilde{l}}/\pi_{\tilde{L}}$, où $\pi_{\tilde{L}}$ correspond à la pression de surface π_s, l'épaisseur entre les niveaux est donnée par $\Delta\sigma_l = \sigma_{\tilde{l}} - \sigma_{\tilde{l}-1}$, où σ_l est calculé par une moyenne géométrique $\sigma_l = \sqrt{\sigma_{\tilde{l}}\sigma_{\tilde{l}-1}}$. Comme pour l'état de base, les valeurs $\sigma_{\tilde{l}}$ sont calculées de sorte que l'état initial soit en équilibre hydrostatique dans l'espace discrétisé verticalement. Autrement dit, pour le profil thermique de l'état de base (*ie :* $\overline{T}_l = \overline{T}(z_l)$), $\sigma_{\tilde{l}}$ sont

1. William Froude (1810-1879) : ingénieur, hydrodynamicien et architecte naval britannique

tels que $\Delta\sigma_l/\sigma_l = \Delta z_l/\overline{H}_l$ avec $\Delta z_l = z_{\tilde{l}-1} - z_{\tilde{l}}$ et $\overline{H}_l = R\overline{T}_l/g$, pour $l \in [1; L]$. Après quelques manipulations algébriques, nous obtenons la relation géométrique suivante :

$$\sigma_{\tilde{l}-1} = \sigma_{\tilde{l}} \left[\sqrt{1 + \left(\frac{\Delta z_l}{2\overline{H}_l} \right)^2} - \frac{\Delta z_l}{2\overline{H}_l} \right]^2, \quad \text{pour} \quad l \in [2; L] \tag{V.2}$$

avec $\sigma_{\tilde{L}} = 1$ et $\sigma_{\tilde{0}} = 0$. Par ailleurs, ce choix de relation pour la coordonnée σ permet d'avoir des niveaux verticaux plus ou moins régulièrement espacés en hauteur pour $\Delta z_l = \Delta z = \text{Cte}$, pour $l \in [2; L]$.

Du fait de l'écriture du système par la coordonnée masse, nous imposons des conditions élastiques au sommet, et matérielles à la surface. Les dimensions du domaine d'intégration diffèrent selon les cas tests étudiés, néanmoins pour les cas sans orographie, des conditions latérales périodiques sont imposées pour la direction horizontale (§2), tandis qu'en présence d'orographie, un rappel vers l'état de base est imposé aux bords latéraux du domaine via une méthode de relaxation de type Davies (1983) [15] (§3). De plus, une zone artificielle d'absorption est introduite sur les premiers niveaux du modèle uniquement pour les cas de forçage orographique (§3) afin de réduire l'impact des réflexions d'ondes dues à l'imposition de la condition élastique au sommet du modèle.

Le système d'Euler dans le plan vertical en coordonnée masse σ est discrétisé spatialement suivant les mêmes méthodes de discrétisation spatiale que celles employées couramment dans le modèle AROME, à savoir : une méthode de transformation spectrale (de Fourier) des champs sur une grille horizontale de collocation (type-A), une méthode des différences finies d'ordre deux sur la grille de Lorenz (type-C), avec une légère modification pour certains opérateurs verticaux dans le cas de la méthode VIPE définie dans le chapitre précédent. Il est à noter que pour tous les schémas examinés ci-dessus, l'advection est Eulérienne, et traitée selon le schéma explicite appliqué par les différents schémas. Pour finir, dans le cas de schémas SI 3-TL et K(3)-Split, pour lesquels un schéma saute-mouton classique est employé pour traiter l'advection, un léger filtre temporel d'Asselin(1972) [4] avec un coefficient de filtrage $\nu = 0{,}05$. Il convient de noter également que dans le cas du schéma HEVI Trap2-Mixed, le traitement VIPE n'est activé que dans les cas tests orographiques, car en absence d'orographie les traitements VIPE et VITE sont équivalents. Les paramètres pertinents des ces différents cas tests seront : le nombre de Courant ondulatoire horizontal $C_* = \overline{c}_s \pi \Delta t/\Delta x$, la vitesse d'advection $\overline{U}$ (qui dans certain cas sera évaluée en nombre de Mach $M_U = \overline{U}/\overline{c}_s$), $S_* = \max|\partial_x z_s|(\Delta x/\Delta z)$ la hauteur maximale de l'orographie et le nombre de Froude $F_r = \overline{U}/(\overline{N}h$ (avec h la hauteur maximale du relief).

2 Expériences sans orographie

Les deux expériences réalisées dans cette partie visent à confirmer les études de stabilité réalisées sans orographie (donc $S_* = 0$ et F_r n'est pas défini). C'est pourquoi, tous les meilleurs schémas-candidats des chapitres précédents sont testés.

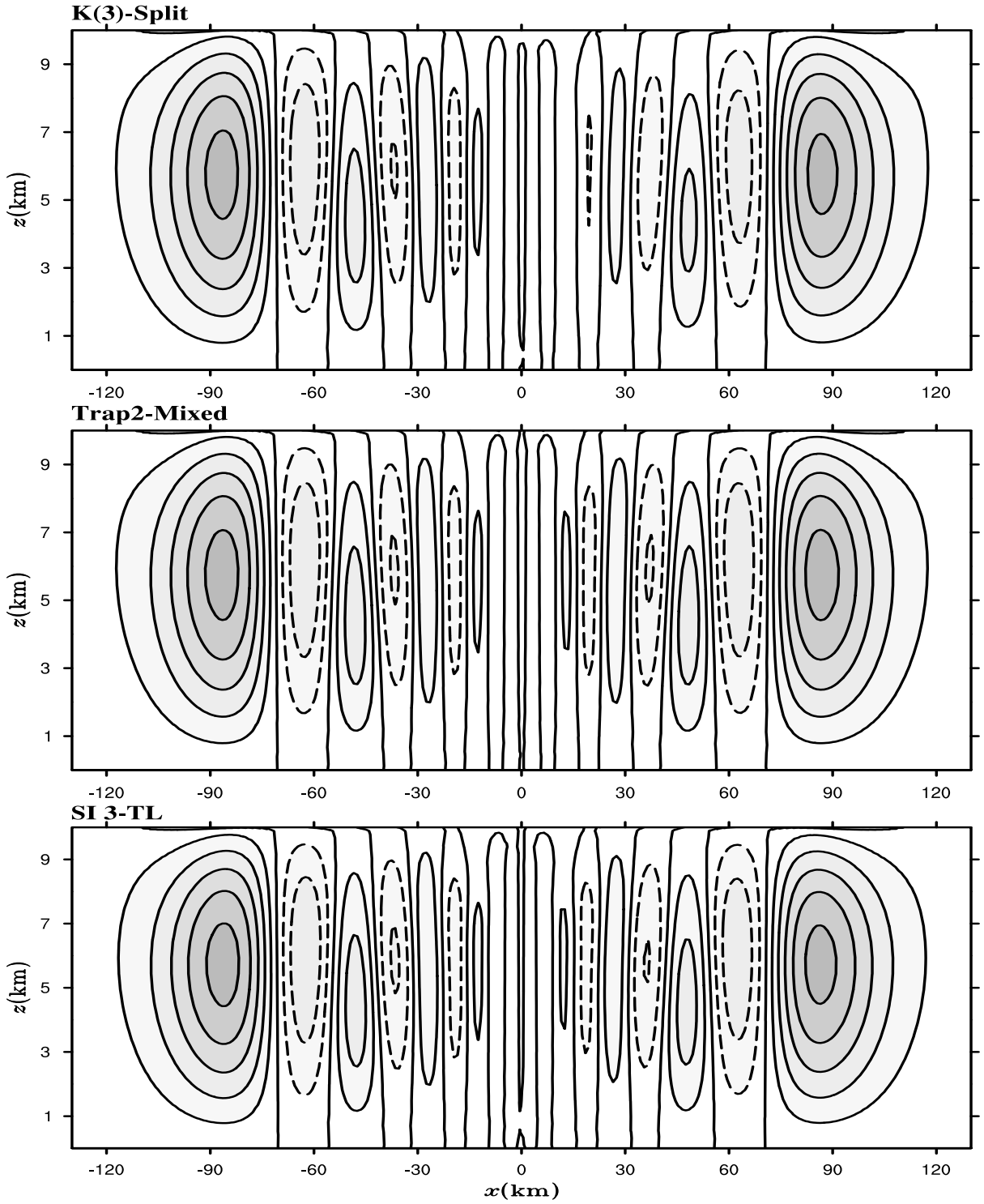


FIGURE V.1 – *Perturbation de température potentielle au bout de* 3000 *s pour les schémas K(3)-Split* ($C_* = 1{,}1$)*, Trap2-Mixed* ($C_* = 1{,}1$)*, et SI 3-TL* ($C_* = 4{,}4$)*.*

Test A : Propagation horizontale des ondes de gravité

Ce test, défini par Skamarock & Klemp (1994) [61], met en évidence la propagation des ondes de gravité dans un écoulement dont le régime est non-hydrostatique. Les ondes sont excitées par une petite perturbation initiale de la température potentielle de la forme d'une bulle chaude :

$$\theta'(x,z) = \Delta\theta_0 \frac{\sin(\pi z/H_T)}{1+(x-x_c)^2/a^2},$$

d'amplitude $\Delta\theta_0 = 0{,}01\,\mathrm{K}$ et qui se superpose à l'état de base ayant une fréquence de Brunt-Väisälä $\overline{N} = 0{,}01\,\mathrm{s}^{-1}$, et un vent horizontal moyen $\overline{U} = 20\,\mathrm{m.s}^{-1}$. Le domaine d'intégration s'étend sur $H_T = 10\,\mathrm{km}$ de haut et 600 km de long. La perturbation est centrée à $x_c = -60\,\mathrm{km}$. Du point de vue de son évolution, cette perturbation tiède de température potentielle est censée irradier dans le deux sens le long de la direction-x pour que finalement son centre soit placé au centre du domaine au bout de 3000 s. La résolution horizontale est de $\Delta x = 1000\,\mathrm{m}$ et la résolution verticale est $\Delta z = 500\mathrm{m}$. Le pas de temps utilisé pour le schéma SI-3TL est de 4 s, et le pas de temps pour des deux méthodes HEVI (Trap2-Mixed, K(3)-Split) est de 1 s.

La Figure V.1 montre les solutions numériques des schémas K(3)-Split, Trap2-Mixed, et SI 3-TL. Nous pouvons constater que, globalement, ces résultats sont en accord avec les autres solutions numériques obtenues pour ce test (Giraldo & Restelli (2008) [26] et Melvin *et al.* (2010) [46]). Par rapport à la solution analytique calculée par Skamarock & Klemp (1994) [61] dans le cadre simplifié du système Boussinesq en coordonnée hauteur (avec des conditions au sommet matérielles). La légère asymétrie de la solution vis-à-vis de la ligne $z = H_T/2$ est due aux effets de compressibilité du système d'Euler. Qualitativement, les résultats semblent proches, avec néanmoins un faible amortissement observé pour le schéma K(3)-Split. Par ailleurs, en mesurant les erreurs quadratiques moyennes (RMS) sur la perturbation de température potentielle, en prenant pour simulation de référence la solution déterminée à partir du schéma explicite RK4 avec un plus petit pas de temps (voir Figure V.2), nous constatons que la schéma Trap2-Mixed est le plus précis, celui dont les RMS sont au plus bas tout au long de l'intégration pour ce cas test. On note également que le schéma Trap2(2,3,2)(-1)-UFpreF produit exactement les mêmes RMS que le schéma Trap2-Mixed.

Afin de mettre en évidence le gain de stabilité apporté par l'utilisation du schéma Trap2-Mixed par rapport à son homologue UFpreF, plusieurs expériences ont été menées qui sont issues de ce test, en faisant varier la vitesse de l'écoulement initial $\overline{U}$. Le protocole expérimental consiste à rechercher de manière empirique les valeurs limites du nombre de Courant C_* pour lesquels le schéma demeure stable pour une valeur du nombre de Mach M_U donnée. Pour $M_U = 0{,}25$ le schéma Trap2-UFpreF admet pour CFL limite $C_* = 1{,}5$ alors que Le Trap2-Mixed autorise jusqu'à $C_* = 1{,}75$. Pour $M_U = 0{,}75$, $C_* = 1{,}1$ pour le cas UFpreF contre $C_* = 1{,}3$ dans le Mixed (pour tous les résultat, *cf* Tableau V.1). Par ailleurs, cette étude nous a permis de confirmer que les traitements UFpreB et UBpreF sont inconditionnellement instables.

Nous avons aussi réalisé la même étude pour le schéma K(M)-Split pour lequel, nous avons étudié l'impact du comportement de la stabilité du schéma pour les fortes advection en fonction du nombre de sous-pas de temps HEVI M. Les résultats sont consignés dans le Tableau V.1 ci-dessous.

Nous constatons que, de manière attendue, le nombre de Courant horizontal limite diminue en fonction de la vitesse d'advection quel que soit le nombre de subdivisions du pas de temps. Ces

M_U	Trap2-Mixed	Trap2-UFPreF	K(1)-Split	K(2)-Split	K(3)-Split
0,25	1,85	1,6	1,17	0,82	0,57
0,50	1,4	1,3	0,92	0,33	0,22
0,75	1,3	1,1	0,54	0,27	0,16

TABLE V.1 – *Nombre de Courant horizontal maximal atteint expérimentalement par les schémas, Trap2-Mixed, Trap2 UFPreF, et K(M)-Split ($M \in \{1;2;3\}$) pour différentes valeurs du nombre de Mach M_U de l'écoulement de base pour l'expérience de la bulle tiède.*

expériences indiquent, en particulier, que le pas de temps critique diminue en fonction de M pour un nombre de Mach donné. Cette preuve confirme que les méthodes aux pas de temps fractionnés sont mal adaptées lorsque la vitesse d'advection est élevée.

Test B : Courant de densité

Ce test non-linéaire, introduit par Straka *et al.* (1993) [64], simule l'évolution d'une perturbation de la température ayant la forme d'une bulle froide définie par :

$$T'(x,z) = \begin{cases} 0, & \text{si} \quad R_d > 1 \\ \\ \Delta T_0 \left[\cos\left(\Pi R_d\right) + 1\right]/2, & \text{si} \quad R_d \leq 1 \end{cases} \tag{V.3}$$

avec :

$$R_d = \sqrt{\left(\frac{x - x_c}{x_d}\right)^2 + \left(\frac{z - z_c}{z_d}\right)^2} \tag{V.4}$$

et $\Delta T_0 = -15\,\mathrm{K}$. Les dimensions de la bulle sont données par : $x_c = 0\,\mathrm{m}$, $x_d = 4000\,\mathrm{m}$, $z_c = 3000\,\mathrm{m}$, et $z_d = 2000\,\mathrm{m}$. La perturbation est placée au centre du domaine d'intégration couvrant 38,4 km sur l'horizontale et 6,45 km sur la verticale. L'atmosphère de base présente une stabilité statique neutre (*ie :* $\overline{N} = 0$) et est supposé au repos ($\overline{U} = 0$).Les résolutions horizontales et verticales de ce test sont prises égales à 75 m. Ce test étant hautement non-linéaire, une légère diffusion numérique est appliquée de manière implicite dans l'espace spectral à la fin de chaque pas de temps. Ce terme, assimilable à une viscosité numérique, est introduit dans les équations pronostiques discrètes sous la forme :

$$\frac{(\hat{\psi}^+_{k,l})^{\mathrm{diff}} - \hat{\psi}^+_{k,l}}{\Delta t} = -\nu_x k^2 \hat{\psi}^+_{k,l} + \frac{\nu_\sigma}{\delta_l}\left[\frac{\pi_{\tilde{l}}}{\pi_{l+1} - \pi_l}\left(\hat{\psi}^+_{k,l+1} - \hat{\psi}^+_{k,l}\right) + \frac{\pi_{\tilde{l}-1}}{\pi_l - \pi_{l-1}}\left(\hat{\psi}^+_{k,l-1} - \hat{\psi}^+_{k,l}\right)\right] \tag{V.5}$$

où $\hat{\psi}^+_{k,l}$ désigne le coefficient de Fourier associé au nombre d'ondes horizontales k d'une variable pronostique quelconque du modèle placée sur les niveaux l et pris à la fin du pas de temps. Des conditions de glissement aux limites supérieure et inférieure domaine sont imposées sur les variables

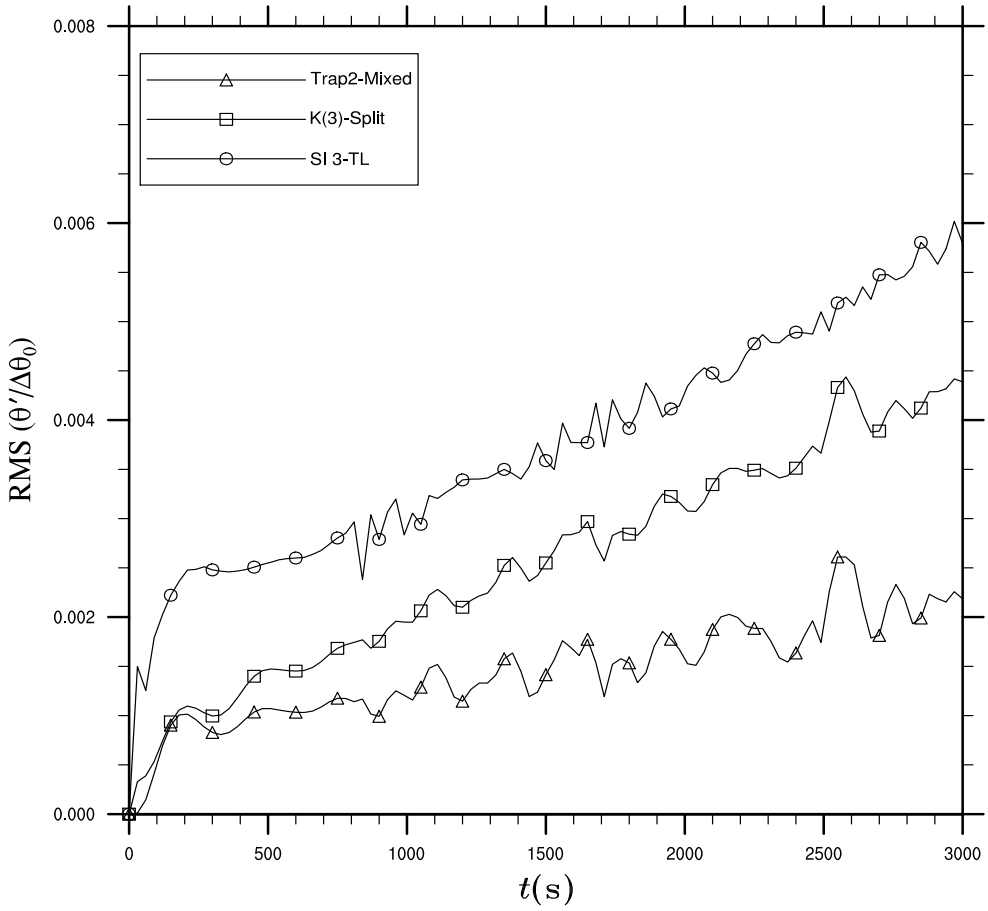


FIGURE V.2 – *Évolution des erreurs quadratiques moyennes des solutions obtenues respectivement via les schémas Trap2-Mixed, K3-Split et le schéma SI 3-TL par rapport à la solution RK4 explicite de référence.*

T et q. Le traitement de la variable w diffère légèrement car elle est placée sur les interfaces. Les coefficients de diffusion sont choisis tels que $\nu_x = 150\,\mathrm{m^2 s^{-1}}$ et $\nu_\sigma = 25 \times 10^{-8}\,\mathrm{m^2 s^{-1}}$ de façon à reproduire des conditions de viscosité similaires à celles du test de Straka *et al.* (1993) [64].

Comme initialement la masse d'air de bulle est plus froide que le reste de l'atmosphère, elle est donc plus dense. C'est pourquoi, elle est censée tomber, s'écraser au sol et générer des cellules tourbillonnaires (appelées *rotors*) induites par le cisaillement vertical du vent. Ceci est caractéristique des instabilités de Kevin-Helmholtz, typiquement mis en évidence dans le cas de ce test et qui sont à l'origine du son comportement hautement non-linéaire. La solution est intégrée pendant une durée de 900 s. La Figure V.3 montre que la solution obtenue à l'aide du schéma Trap2-Mixed ($C_* = 0{,}75$) reproduit correctement l'évolution de la perturbation de température potentielle.

Ce test a également été reproduit avec les autres schémas : SI 3-TL ($C_* = 3$), K(3)-Split ($C_* = 0{,}75$), et RK4 ($C_* = 0{,}15$). Les solutions associées à ces trois derniers schémas n'ont pas été affichées ici compte tenu de leur grande ressemblance visuelle avec la solution obtenue avec la méthode Trap2-Mixed. Toutefois, la Figure V.4 exhibe une coupe horizontale dans la direction des x positifs en $z = 1{,}2\,\mathrm{km}$ pour les solutions respectives des quatre schémas à l'instant 900 s. On

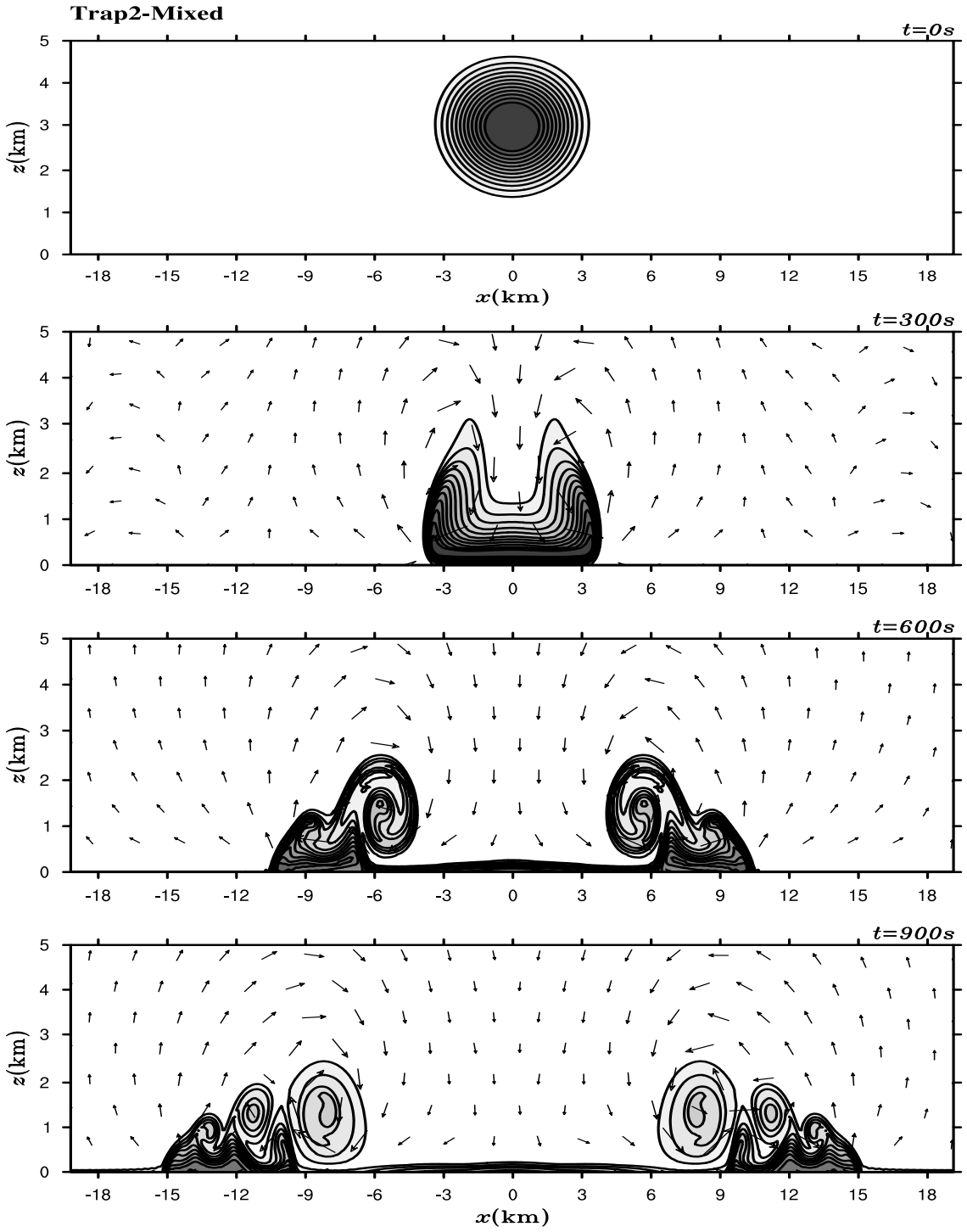


FIGURE V.3 – *Évolution de la perturbation de la température potentielle de l'état initial* $t = 0\,\mathrm{s}$ *à* $t = 900\,\mathrm{s}$ *pour le schéma Trap2(2,3,2)-Mixed proposé. Les iso-lignes sont placées tous les* $1\,\mathrm{K}$*. Le champ de vecteur correspond à la circulation du vent dont la vitesse maximale se situe autour de* $35\,\mathrm{m.s^{-1}}$*.*

y remarque que toutes les solutions sont en avance par rapport à la solution RK4, considérée ici comme la solution de référence. En effet, pour les deux schémas HEVI et pour le schéma SI, les positions des trois rotors (indiquées par les trois minima relatifs) ainsi que la localisation du front avant de la perturbation sont toutes légèrement en avance par rapport aux positions obtenues via le schéma RK4. Il est à noter également que l'amplitude minimale de température potentielle au centre du premier rotor est sur-estimée approximativement $-0{,}3\,\mathrm{K}$ par tous les schémas utilisés.

Ce test ne permet pas de déterminer quel est le schéma HEVI le plus précis. En revanche, du fait de l'introduction du terme diffusif, et du caractère hautement non-linéaire de cette expérience,de nouvelles conditions de stabilité viennent renforcer la contrainte CFL linéaire étudiée dans les chapitres précédents. Il apparaît que la version UFpreF du schéma Trap2(2,3,2)(-1) ne permet pas d'obtenir des nombres de Courant plus grands que 0,15. Cette expérience plaide donc, une nouvelle fois, pour l'utilisation des méthodes RK-IMEX HEVI à quatre tableaux de Butcher qui, par leurs traitements davantage implicites que les versions UFpreF, semblent également plus stables.

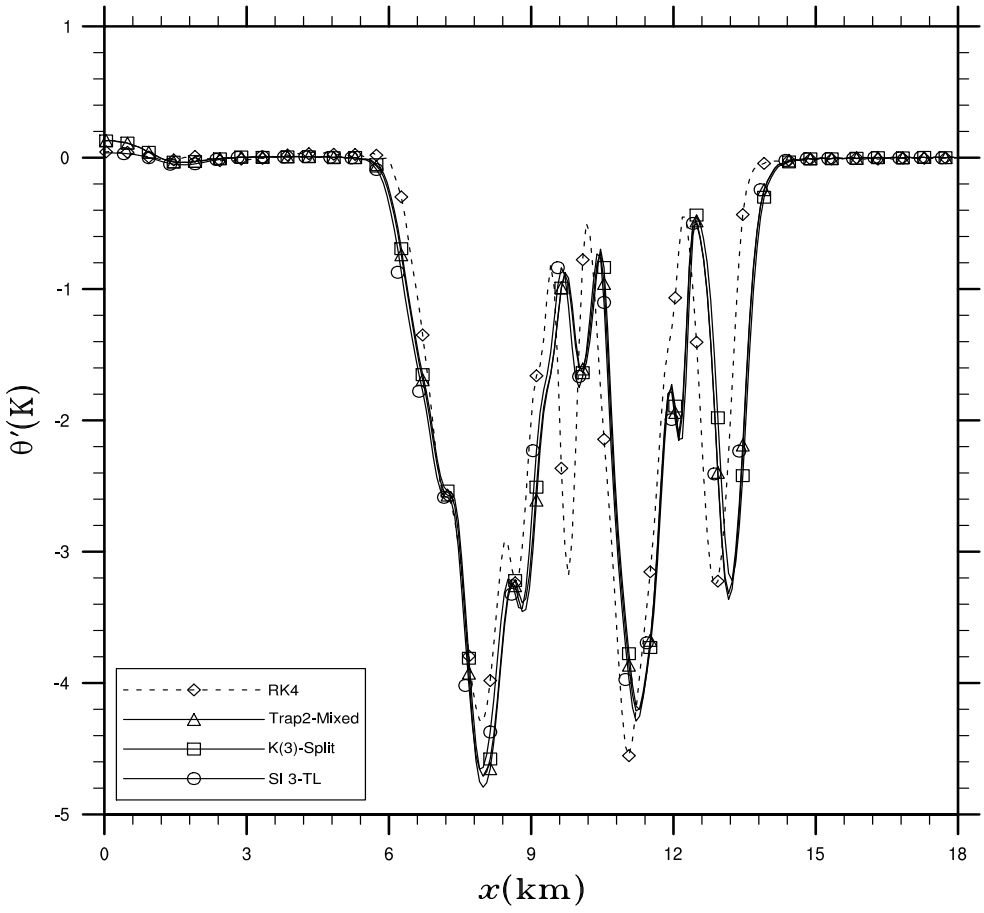


FIGURE V.4 – *Coupe horizontale en $z = 1{,}2\,km$, de la perturbation de température potentielle à $900\,s$ pour les schémas RK4 (référence), Trap2-Mixed, K3-Split, et SI 3-TL.*

Ces premières études nous ont donc permis de mettre en lumière le fait que les schémas fractionnant le pas de temps semblent inappropriés pour simuler des écoulements avec une forte vitesse.

De plus, nous avons confirmé que les schémas Mixed sont qualitativement comparables à d'autres schémas d'ordre 2 et que le schéma Trap2-Mixed est, non-seulement plus stable pour des écoulements proches d'un cas linéaire, mais aussi, bien plus stable pour des écoulements plus réalistes. Reste maintenant à étudier l'impact d'un forçage orographique sur la stabilité des schémas HEVI appliqués au système d'Euler.

3 Écoulements orographiques

Les cas tests présentés dans cette partie sont issus de la littérature autour des écoulements forcés par la présence d'une orographie. La forme de cette orographie et sa hauteur maximale influent sur la forme de la réponse atmosphérique et sur la nature d'écoulement. Dans ce qui suit, deux cas tests pour deux formes de relief différents sont présentés, l'objectif étant de montrer que, d'une part, l'implémentation du schéma HEVI Trap2-Mixed permet de reproduire fidèlement l'évolution de l'écoulement au-dessus de ces reliefs, et d'autre part, montrer que la version VIPE est plus stable que la version VITE. Nos résultats seront comparés à ceux des schéma SI, et RK4 implémentés également ici en coordonnée masse pour faciliter les comparaisons, ainsi qu'aux autres solutions obtenues en coordonnée hauteur z présentées dans la littérature (Melvin *et al.* (2010) [46], Simarro & Hortal (2012) [58]).

Test C : Quasi-linéaire non-hydrostatique

Ces premiers tests, avec une faible orographie, visent à évaluer qualitativement la réponse des schémas que nous proposons. Pour cela, nous étudions deux types de reliefs différents.

Cas d'un relief de type Bubnová :
Ce test, proposé par Bubnová *et al.* (1995) [9], cherche à simuler l'écoulement au dessus de relief de la forme d'une cloche de type *Agnesi* définie par :

$$z_s(x) = h\frac{a^2}{x^2 + a^2} \tag{V.6}$$

avec une hauteur maximale h du relief de 100 m, et une demi-largeur $a = 5\Delta x$. Les autres paramètres de cette expérience sont : $\overline{U} = 15\,\mathrm{m.s^{-1}}$ et $\overline{N} = 0{,}02\,\mathrm{s^{-1}}$. Le domaine d'intégration couvre 50 km et 20 km de haut avec une éponge placée à partir de 14 km permettant d'assurer la non-réflectivité du bord supérieur du domaine. Aucune diffusion numérique n'est appliquée pour ce test. Les résolutions horizontale et verticale sont telles que $\Delta x = \Delta z = 100\,\mathrm{m}$. Ainsi, le nombre de Froude de l'écoulement de base est défini par $F_r = 7{,}5$. Dès que le nombre de Froude est supérieur à 1, alors la longueur d'onde de l'écoulement est plus grande que celle de la barrière. Il y a par conséquent une accélération du vent au sommet et, après une distance d'environ trois fois la longueur de l'obstacle, l'écoulement retrouve ses caractéristiques initiales.

La Figure V.5 montre les solutions, en terme de perturbation, de vent horizontal $u - \overline{U}$ et vertical w au bout de 3000 s d'intégration pour les schémas Trap2-Mixed ($C_* = 1{,}75$), RK4 ($C_* = 0{,}5$) et SI 3-TL ($C_* = 11$). De manière générale, la méthode SI 3-TL et Trap2-Mixed semblent fidèlement représenter ces ondes orographiques du fait que les positions des perturbations ainsi que leurs

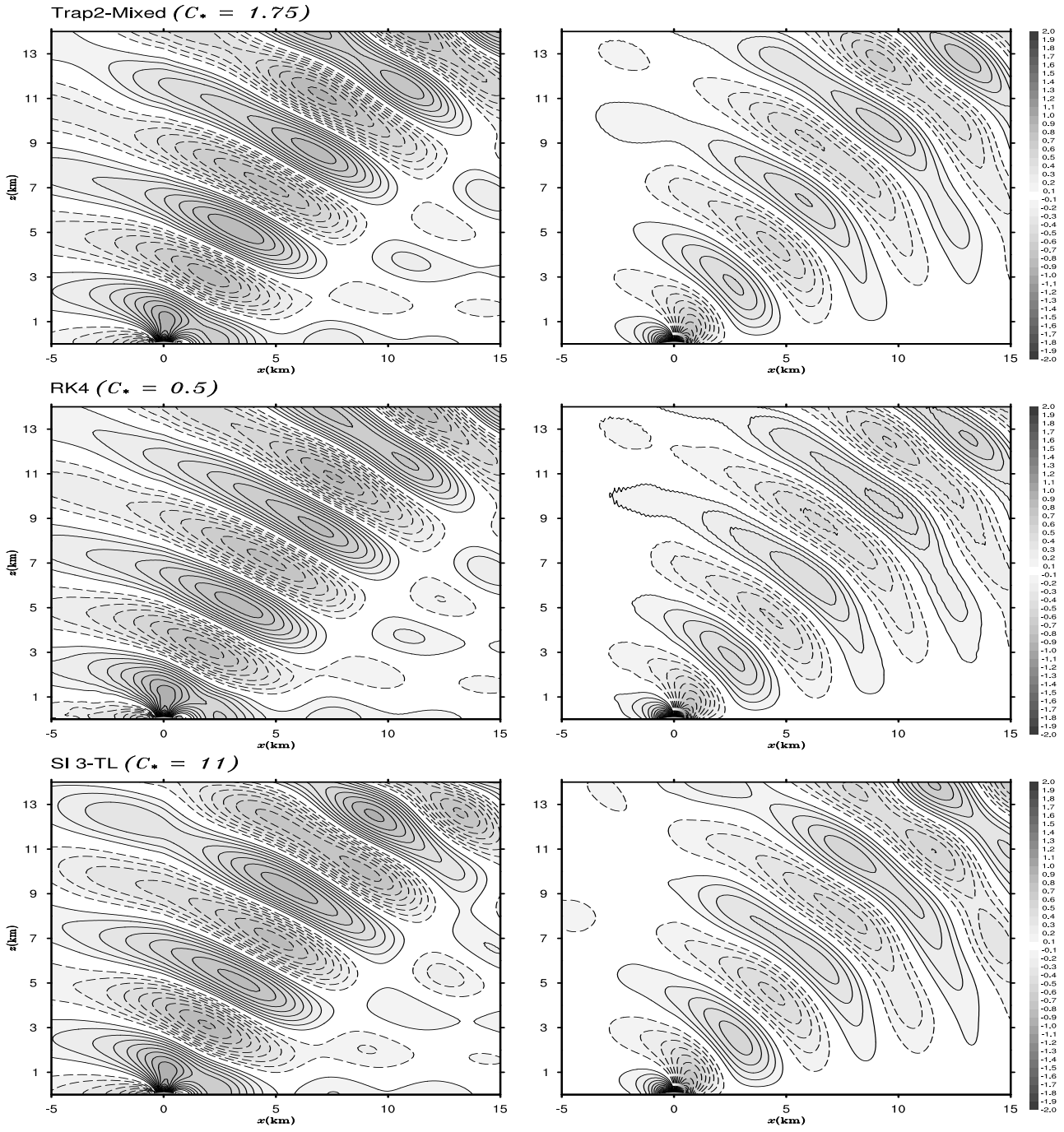


FIGURE V.5 – *Vitesse de la perturbation de vent horizontal $u' = u - \overline{U}$ (à gauche) et du vent vertical w (en* $\mathrm{m.s^{-1}}$*) au bout de* 3000 s *d'intégration pour l'expérience de Bubnová et al. (1995) [9]. En haut, le schéma Trap2-Mixed, au milieu le schéma RK4 et en bas le SI 3-TL. Les isolignes sont tracées tous les* $0{,}1\,\mathrm{m.s^{-1}}$*, les isolignes tiretées correspondent aux valeurs négatives.*

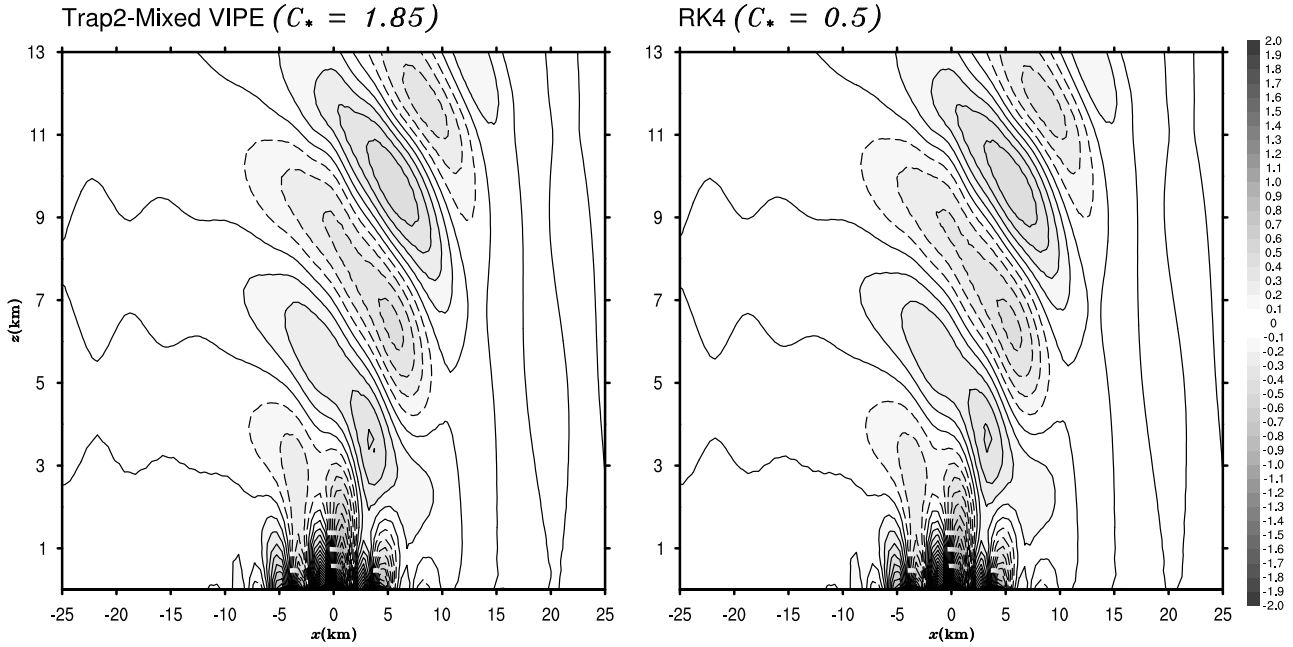


FIGURE V.6 – *Vitesse du vent verticale w (en $\mathrm{m.s^{-1}}$) au bout de 12 h d'intégration pour l'expérience de Schär et al. (2002) [57]. À gauche, le schéma Trap2-Mixed, et à droite, le schéma RK4. Les isolignes sont tracées tous les $0{,}05\,\mathrm{m.s^{-1}}$, les isolignes tiretées correspondent aux valeurs négatives.*

amplitudes globales sont similaires. Les plus grosses différences entre ces schémas apparaissent au delà de 9 km d'altitude. À ces niveau là, le SI 3-TL semble à la fois amortir le signal et légèrement déplacer l'amplitude maximale de l'onde. À l'inverse, pour les niveaux inférieurs, il semble que le schéma SI surestime l'amplitude du signal. Pour ce cas-là, il semble visiblement que le schéma Trap2-Mixed soit plus proche de la solution RK4 de référence.

Cas d'un relief de type Schär *:*
Pour ce cas test orographique, on considère un relief de forme un peu plus complexe donnée par Schär *et al.* (2002) [57] :

$$z_s(x) = h \exp\left[-(x/a)^2\right] \cos^2\left[\pi(x/b)\right] \tag{V.7}$$

avec $h = 250\,\mathrm{m}$, $a = 5000\,\mathrm{m}$, et $b = 4000\,\mathrm{m}$, et cette fois les paramètres de l'état de base suivants : $\overline{N} = 0{,}01\,\mathrm{s^{-1}}$, $\overline{U} = 10\,\mathrm{m.s^{-1}}$, et $T_s = 288\,\mathrm{K}$. Les mailles sont respectivement $\Delta x = 500\,\mathrm{m}$ et $\Delta z = 312\,\mathrm{m}$. Ainsi, le paramètre $S_* = 0{,}32$ n'est pas suffisamment élevé pour remarquer une différence importante, en terme de stabilité, entre le Trap2(2,3,2)(-1) UFpreF VITE et le Trap2-Mixed VIPE. Comme de plus les résultats sont qualitativement équivalents, nous n'illustrons que le schéma Trap2-Mixed VIPE.

Ce choix de topographie du relief met en scène à la fois des perturbations de grandes échelles associées à sa partie gaussienne et des perturbations de plus petites échelles plutôt reliées à sa modulation sinusoïdale. Ainsi, les ondes de gravité forcés par ce relief présentent deux composantes spectrales dominantes : l'une de grande échelle, plutôt de nature hydrostatique, caractérisée par la

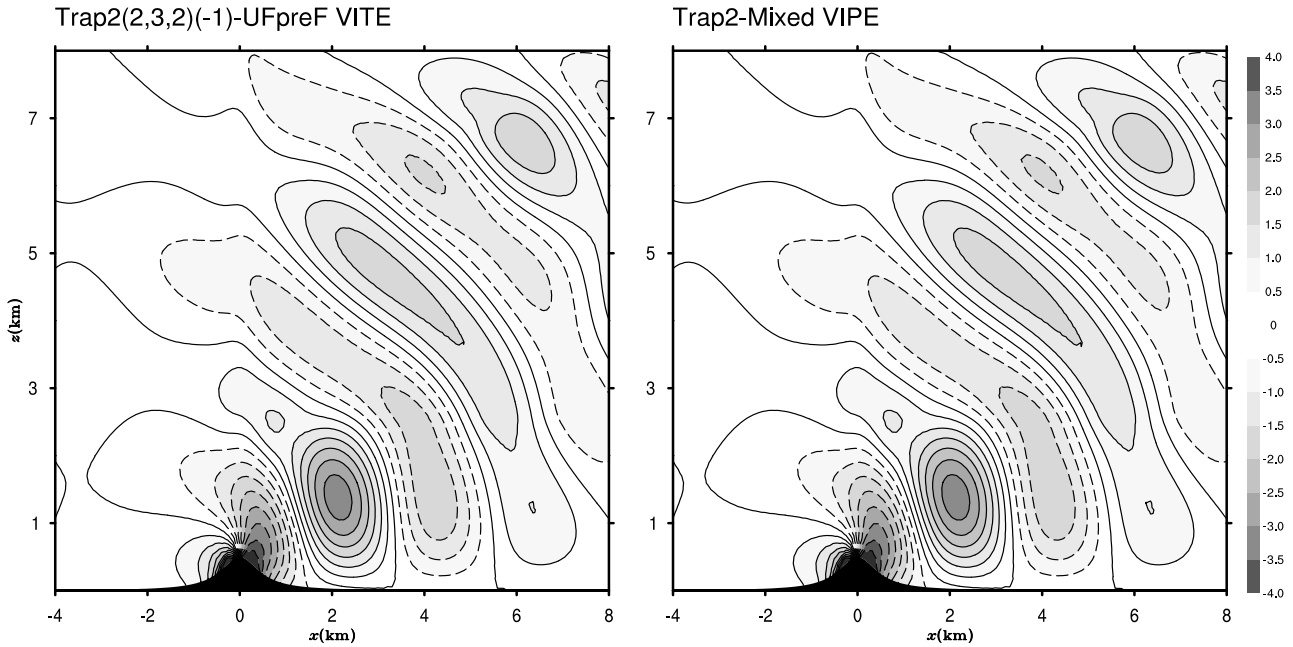


FIGURE V.7 – *Vitesse du vent verticale w (en $\mathrm{m.s^{-1}}$) au bout de 4000 s d'intégration pour l'expérience de Budnová et al. (1995) [9]. À gauche, le schéma Trap2(2,3,2)(-1) VITE, et à droite, le schéma Trap2-Mixed VIPE. Les isolignes sont tracées tous les $0{,}2\,\mathrm{m.s^{-1}}$, les isolignes tiretées correspondent aux valeurs négatives.*

propagation verticale de la perturbation très haut dans l'atmosphère, et l'autre de petite échelle générée par les variations sinusoïdales de la topographie mais dont l'amplitude décroît très vite avec la hauteur due aux effets non-hydrostatiques.

Les résultats de ces expériences, illustrées à la Figure V.6, s'identifient particulièrement bien avec les résultats illustrés à la Figure 9 de Simarro & Hortal (2012) [58] pour lequel ils testent un traitement Eulérien de l'advection et une discrétisation verticale éléments finis avec un schéma SI. Ainsi, cette expérience montre, une fois de plus, le bon comportement qualitatif de schéma Trap2-Mixed, dont nous avions démontré son ordre de précision.

Test D : Non-linéaire non-hydrostatique

Ces dernières batteries d'expériences visent à confirmer deux points liés au traitement de l'orographie. Le premier est de s'assurer que, malgré le traitement implicite VIPE, il n'y a pas d'amortissement (et donc de dégradation) du signal par rapport à un traitement explicite VITE. Le second point est de confirmer qu'il existe des orographies pour lesquelles la version VITE est moins stable que la version VIPE.

Pour ces cas fortement non-linéaires, nous avons introduit une légère diffusion horizontale spectrale implicite qui amortit de 1% l'amplitude de l'onde $2\Delta x$ toutes le cinq itérations. En comparaison, la diffusion utilisée par Straka *et al.* (1993) [64] amortit de 5% cette même onde tous les pas de temps.

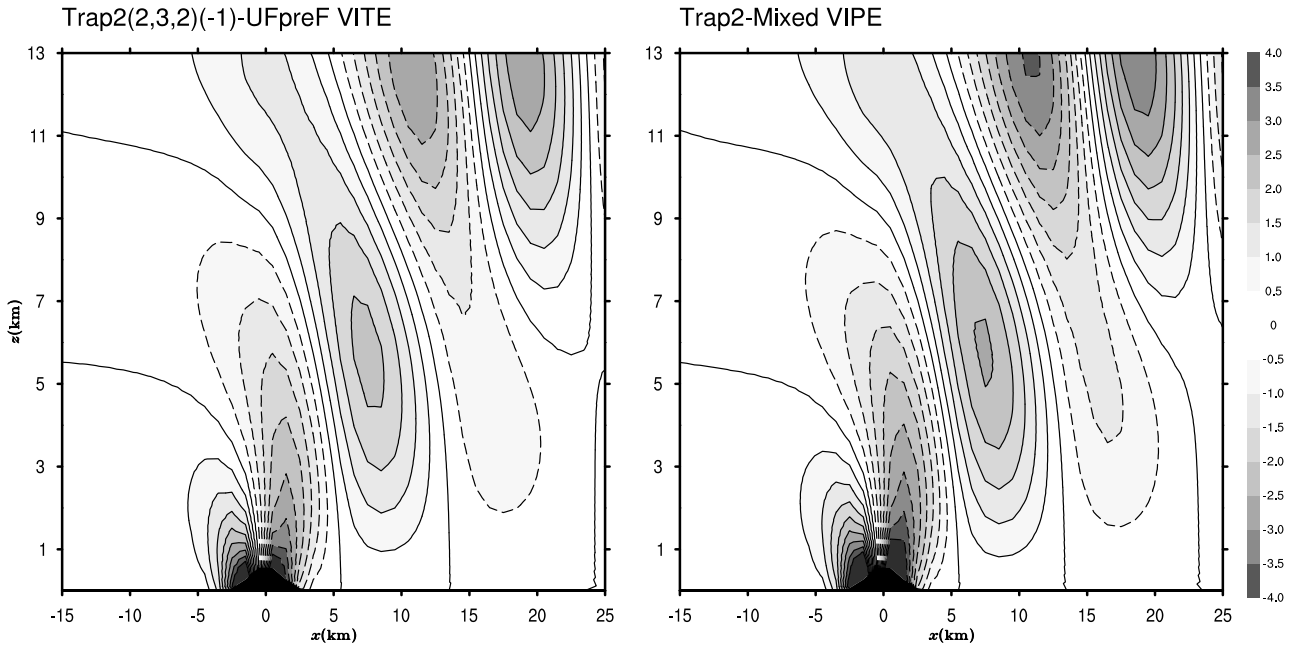


FIGURE V.8 – *Vitesse du vent verticale w (en $\mathrm{m.s^{-1}}$) au bout de 6 h d'intégration pour l'expérience inspirée de Zängl (2012) [74]. À gauche, le schéma Trap2(2,3,2)(-1) VITE avec $h = 500\,\mathrm{m}$, et à droite, le schéma Trap2-Mixed VIPE avec $h = 700\,\mathrm{m}$. Les isolignes sont tracées tous les $0{,}5\,\mathrm{m.s^{-1}}$, les isolignes tiretées correspondent aux valeurs négatives.*

Cas d'un relief de type Agnesi :

La première expérience s'inspire du cas non-linéaire non-hydrostatique défini par Budnová *et al.* (1995) [9]. L'orographie a la forme d'un relief Agnesi pour lequel $h = 500\,\mathrm{m}$, $a = 500\,\mathrm{m}$ et $\overline{U} = 5\,\mathrm{m}.s^{-1}$. Le profil vertical de la température est défini avec $\overline{N} = 0{,}01\,\mathrm{s^{-1}}$ et $T_s = 285\,\mathrm{K}$. Ainsi, le nombre de Froude est égal à 1. Les paramètres numériques sont $\Delta x = \Delta z = 100\,\mathrm{m}$ et la durée d'intégration est de 8000 s avec un pas de temps tel que $C_* = 1{,}65$ pour les deux schémas.

Les résultats de cette expérience sont les deux illustrations de la Figure V.7 pour le vent vertical w. À gauche, nous avons les résultats de l'intégration pour le schéma Trap2(2,3,2)(-1) UFpreF avec un traitement VITE, et à droite le résultat de cette même expérience pour le schéma Trap2-Mixed VIPE. La proximité des résultats confirme d'une part que le traitement implicite des termes orographique n'altère pas le signal par rapport à un traitement explicite classique, et d'autre part que le schéma Trap2-Mixed est bien un schéma d'ordre 2. La seule différence entre ces deux schémas est le nombre moyen d'itérations nécessaires pour résoudre la quasi-Newton. Alors que sur l'ensemble de l'intégration, le nombre d'itérations par pas de temps est de 3 pour le schéma employant la méthode VIPE, elle n'est que de 2,33 pour le schéma VITE. Ceci s'explique par le fait que le problème VIPE reste plus complexe à résoudre, et nécessite donc plus d'itérations. Mais ce surcoût est uniquement présent au début de l'expérience car dès que la durée d'intégration dépasse 2000 s, alors les schémas convergent tous deux en seulement 2 itérations par pas de temps. Ainsi, mécaniquement, pour une durée d'expérience plus longue, ces moyennes tendent à se confondre.

Cas d'un relief de type Gaussienne :
Le but de cette expérience est de vérifier que le traitement VIPE permet, en utilisant un même pas de temps, d'accepter des pentes plus élevées qu'un traitement explicite classique. Pour cela, nous utilisons un relief comme défini par Zängl (2012) [74] :

$$z_s(x) = h \exp\left[-(x/a)^2\right] \tag{V.8}$$

avec $a = 1665\,\mathrm{m}$. Le profil de la température est rigoureusement le même que dans l'expérience avec la montagne de Schär *et al.* (2002) [57], à savoir que $\overline{N} = 0{,}01\,\mathrm{s}^{-1}$, et $T_s = 288\,\mathrm{K}$. Seule la vitesse d'advection initiale change pour $\overline{U} = 20\,\mathrm{m.s}^{-1}$. Afin de montrer l'importance du rapport d'aspect pour les schémas HEVI, nous utilisons des mailles définies par $\Delta x = 1000\,\mathrm{m}$ et $\Delta z = 100\,\mathrm{m}$. La durée d'intégration est de 6 h.

Pour mettre en évidence la meilleure stabilité du traitement VIPE, nous fixons, pour les deux schémas, le même pas de temps $\Delta t = 1{,}4\,\mathrm{s}$ (soit $C_* \approx 1{,}6$) et nous cherchons expérimentalement la hauteur maximale h de la gaussienne, pour que le schéma soit stable. Avec les paramètres que nous avons fixés, nous obtenons dans le cas VITE $h = 500\,\mathrm{m}$ (ce qui correspond à $S_* \approx 2{,}34$), alors que pour le cas VIPE, nous avons pu atteindre $h = 700\,\mathrm{m}$ (soit $S_* \approx 3{,}29$). La Figure V.8 montre le résultat de ces deux expériences. Il faut noter que l'on observe le même comportement sur le nombre d'itérations, pour résoudre le quasi-Newton, que dans les expériences précédentes, à savoir pour les itérations au-delà d'un certain temps d'intégration (ici $t = 8000\,\mathrm{s}$), alors les deux méthodes résolvent leurs problèmes non-linéaires en deux itérations.

Cette petite recherche de hauteur maximale, autorisée par la stabilité des schémas VITE ou VIPE, tendent à valider nos intuitions issues des analyses du chapitre précédent, à savoir que le traitement implicite des termes orographiques améliore la stabilité des schémas, tout en maintenant leurs précisions, par rapport au traitement VITE. Néanmoins, comme nous l'avions déjà suggéré au chapitre précédent, il semble que les analyses dans le cas linéaire soient très optimistes, et que les effets d'une orographie plus réaliste imposent une contrainte plus forte sur le pas de temps critique.

4 Synthèse des résultats expérimentaux

Cette partie expérimentale a permis de valider plusieurs résultats importants :

- Les schémas HEVI étudiés ici sont qualitativement proches, particulièrement ceux des méthodes RK-IMEX d'ordre 2, que cela soit avec deux ou quatre tableaux de Butcher.
- La méthode K(M)-Split devient trop inefficace avec un nombre de découpes $M \geq 1$ pour une vitesse d'advection importante ($M_U \geq 0.5$).
- Le traitement implicite des termes orographiques permet de gagner en stabilité en présence de fortes pentes, particulièrement dans le cas où le rapport d'aspect est grand ($\Delta x/\Delta z \geq 10$)

Ces expériences confirment plusieurs de nos intuitions. La première est que les méthodes au pas de temps fractionnés semblent trop peu stables pour être efficaces dans le cas d'un modèle opérationnel (pour lequel certain jets peuvent atteindre des vitesses comparables à celles du son). La seconde est qu'il est possible d'introduire plusieurs autres schémas Runge-Kutta pour intégrer les équations d'Euler sans impacter la qualité de l'intégration, tant que les conditions que nous avons établies sont respectées. La dernière semble indiquer que le traitement VIPE augmente encore la stabilité en présence d'orographie.

Conclusion et perspectives

Ce travail de thèse visait à élaborer le schéma HEVI , permettant l'utilisation du plus grand pas de temps critique possible, pour intégrer le système d'Euler pleinement compressible en coordonnée masse. Dans ce travail de recherche, deux approches de type HEVI ont été étudiées : les méthodes aux pas de temps fractionnés, conçues pour être le plus économique possible, qui utilisent deux pas de temps distincts pour traiter les processus lents et les processus rapides, et les méthodes Runge-Kutta IMEX assurant l'intégration de l'ensemble des processus sur le même pas de temps via des méthodes s'opérant sur plusieurs étapes.

Dans le cas de l'approche du pas de temps fractionné, alors que les termes linéaires sont traités par un schéma de type forward-backward pour les termes d'ajustements horizontaux et un schéma implicite pour les termes d'ajustements verticaux, les modèles opérationnels utilisant cette méthode traitent les termes responsables de la propagation des processus considérés comme plus lents soit par un schéma saute-mouton, soit par un schéma Runge-Kutta-3. Nous avons revisité le schéma Kurihara (prédicteur/correcteur saute-mouton/trapézoïdal) pour traiter ces processus lents, et nous avons mis en évidence qu'il suffisait d'appliquer un simple filtre temporel de type Asselin, d'usage commun en PNT, pour obtenir un gain substantiel en stabilité du schéma, ainsi qu'une meilleure précision par rapport au schéma saute-mouton. Ce nouveau schéma, que nous avons nommé K(M)-Split, où M, est le nombre de découpes du grand pas de temps, présente une meilleure efficacité par rapport aux autres schémas de sa catégorie, notamment ceux utilisés en opérationnel. Mais les tests numériques ont prouvé la faible résistance de la stabilité de ces schémas en présence de forte advection, ce qui montre qu'il est plus efficace de n'avoir qu'un seul pas de temps pour intégrer l'ensemble des processus.

Dans le cas des méthodes RK-IMEX, nous avons montré que le traitement forward-backward des termes d'ajustements horizontaux, entraîne une augmentation significative de la stabilité uniquement en absence d'advection. Toutefois, en présence celle-ci, le traitement implicite de ces termes d'ajustements provoque une forte diminution de la stabilité de ces schémas, allant jusqu'à l'inconditionnelle instabilité pour la méthode Trap2(2,3,2)(-1). Pour pallier cette difficulté, nous avons démontré la faisabilité d'introduire plusieurs autres schémas Runge-Kutta pour traiter l'ensemble des termes, tout en gardant un schéma d'ordre deux. Même si, *a priori*, il est possible d'appliquer à chaque terme son propre schéma d'intégration temporelle, nous avons mis en évidence que certains termes devaient être intégrés de la même façon pour retrouver une équation de structure discrète semblable à celle du problème linéaire continu (et ainsi assurer que les ondes rapides soient proches de la solution physique). Cette étude a permis de savoir qu'il était suffisant d'avoir six schémas différents, dont un qui soit purement explicite et un, au moins, ayant au mi-

nimum une itération implicite. Grâce à cela, nous avons pu créer une nouvelle classe de schémas à la fois plus stable que celle déjà étudiée dans la littérature, et qui n'a aucun coût numérique supplémentaire. De plus, ces schémas évitent un autre écueil, celui d'annuler la vitesse de groupes des ondes numériques. Nous nous sommes restreint à n'utiliser que quatre schémas Runge-Kutta différents en intégrant les termes X et Y soit par les schémas intégrant respectivement la divergence horizontale et le gradient horizontal de pression (traitement VITE), soit de traiter ces deux termes de manière implicite avec le même schéma intégrant la partie verticale (traitement VIPE). Il faut noter que ces deux traitements peuvent s'appliquer à n'importe quel schéma HEVI, y compris pour les méthodes aux pas de temps fractionnés, examinées dans ces travaux, mais aussi pour toutes coordonnées suivant le terrain. Néanmoins, ce traitement ne peut permettre de résoudre complètement les problèmes engendrés par l'orographie. Nous avons montré, et vérifié expérimentalement, que ce nouveau traitement VIPE permettait une bien meilleure stabilité du schéma en présence de fortes pentes lorsque des méthodes spectrales sur l'horizontale étaient appliquées.

En somme, l'ensemble des études menées dans le cadre de cette recherche tend à indiquer que le meilleur schéma HEVI que nous avons pu élaborer (*ie :* celui qui soit le plus apte à être utilisé en opérationnel) soit le schéma Trap2-Mixed avec un traitement VIPE. Le fait que ce candidat soit tout aussi précis, mais également plus le plus stable des autres candidats examinés ici, que ce soit dans le cadre linéaire ou dans les expériences académiques, le rend, *a priori*, plus efficace pour un vrai modèle opérationnel. Mais il convient d'être prudent, et seules des études plus poussées, avec un vrai modèle, pourraient convertir cette conjecture en certitude. En effet, de nombreux autres termes sont encore à inclure à nos modèles, en particulier ceux de la physique. Où la physique doit-elle être appelée, lors du calcul des états intermédiaires, afin de garantir la meilleure stabilité possible du schéma ? Le traitement de ces nouveaux termes peut-il encore infirmer les conjectures sur le meilleur candidat ?

Par ailleurs, d'autres pistes d'amélioration sont encore à exploiter. Pour cela, il est nécessaire de remettre en cause certains paramètres fondamentaux de ce travail. Nous avons toujours utilisé les mêmes variables pronostiques qui sont celles du modèle d'AROME. Or, il existe certains jeux de variables qui permettent de déplacer formellement certains termes d'une équation à l'autre, notamment les termes liés à la présence d'orographie. Par exemple, il est possible d'utiliser la variable de température potentielle et du géopotentiel en lieu et place de la température et de la pression (Klemp (2007) [34]). Cette forme du système d'Euler ne contient pas le terme croisé X du fait que la divergence 3D est absente des équations pronostiques de ces deux variables. Pour supprimer ces termes orographiques, il est aussi possible d'utiliser les variables covariantes (Bénard *et al.* (2005) [8], Simarro & Hortal (2012) [58], Dubos & Tort (2014) [16]). Ainsi, bien qu'il soit clair que la stabilité linéaire soit indépendante de la forme du système, il est peut-être encore possible d'augmenter la stabilité de ces schémas par la suppression de contraintes non-linéaires, notamment en présence de fortes pentes. L'étude doit donc être étendue pour le système pleinement compressible écrit pour ces autres jeux de variables pronostiques.

Une autre piste d'amélioration de la stabilité concerne le traitement de l'advection. Les études tout au long de ces travaux ont été réalisées avec un traitement Eulérien de ce processus pour garantir une certaine conservation de la masse. Pour chaque schéma HEVI, nous avons remarqué que le traitement explicite de ce processus entrainait toujours une diminution plus ou moins importante de la stabilité, dont nous n'avons pas clairement établi les causes dans le cas des schémas

RK-IMEX HEVI. Ainsi, ne serait-il pas envisageable de recourir à un traitement Lagrangien de ce processus afin de diminuer la contrainte sur la stabilité ? Comme il est absurde de parler de prix sans savoir ce qu'il y a à acquérir, il faut avant tout regarder s'il y a un gain de stabilité en utilisant ces méthodes et de combien, avant d'évaluer le coût du traitement Lagrangien. Ce type de traitement n'a encore jamais été envisagé dans un contexte HEVI, alors que les petits pas de temps imposés par la contrainte HEVI ne provoquent pas une grande perte de scalabilité de l'algorithme lors de la recherche du point d'origine. Ainsi, il est peut-être envisageable d'utiliser le schéma ARK2(2,3,2) en version UFpreB (qui est le plus stable des schémas RK-IMEX au repos) avec un traitement lagrangien afin de pouvoir obtenir un nombre de Courant critique proches de $2\sqrt{2}$. Par ailleurs, le cas Trap2-Mixed est un schéma de type prédicteur/correcteur pour lequel ce traitement lagrangien se combine parfaitement bien (Cullen (2001) [13]).

Nous savons, par ailleurs, que la discrétisation spatiale a une importance capitale sur la stabilité des schémas (ne serait-ce que sur l'écriture des nombres de Courant et donc de la contrainte CFL). Comme nous n'avons voulu étudier uniquement que la qualité et la stabilité de ces schémas temporels, nous avons utilisé des méthodes spectrales pour l'horizontale. Ceci a permis de s'affranchir des questions sur la consistance des traitements de termes X et Y lorsqu'une coordonnée suivant le terrain est utilisée. Or, comment représenter fidèlement ces termes dans le cas où des schémas de types différences finies ou volumes finis sont appliqués, et quel serait l'impact de ces discrétisations ? Dans ce cas-là, il semble important de prendre soin de traiter de manière consistante ces termes orographiques aux abords du relief, en s'appuyant sur des méthodes d'extrapolation, proposées par Zängl (2012) [74] ou Mahrer (1984) [43] pour étendre le pas de temps critique.

Pour terminer, il faut aussi soulever le fait que les architectures des futurs calculateurs sont fortement dépendants des constructeurs, et que l'on peut encore envisager la tendance vers des machines à mémoire de plus en plus distribuée soit corrigée. Dans ce contexte, il est encore possible que les méthodes SI demeurent les plus efficaces.

Annexe A

Version originale de l'article soumis au *Q. J. R. Meteor. Soc.*

Récemment, plusieurs schémas HEVI, basés sur une approche IMEX avec des méthodes en plusieurs étapes, ont été développés pour résoudre les problèmes liés à la scalabilité des algorithmes. Chacun de ces schémas procède des caractéristiques qui lui sont propres et qui le rend, *a priori*, envisageable pour un modèle opérationnel. Cet article compare trois RK-IMEX HEVI dont l'étude de stabilité est réalisée sur le système d'Euler pleinement compressible linéarisé, qui supporte l'ensemble des ondes rapides observées en PNT (acoustiques et gravité). Chaque schéma est analysé avec les deux versions UFpreF et UFpreB. La propagation des ondes de gravité se trouve être généralement bien représenté quelle que soit la version, et la version UFpreB est toujours plus stable pour ce système au repos. Or, l'introduction d'un écoulement constant rend plus ou moins instables toutes les variantes UFpreB. Malgré le fait que l'origine de ces instabilités soit encore mal définie, une étude dans le cadre unidimensionnel a permis la création d'une nouvelle classe de schémas pour contourner ce problème : les schémas RK-IMEX avec quatre tableaux de Butcher. Les deux tableaux de Butcher supplémentaires, exclusivement réservés au traitement des termes d'ajustements horizontaux, a permis de gagner en stabilité par rapport aux précédents schémas étudiés dans la littérature. Nous montrons que ces schémas sont aussi précis et à coût constant. Un schéma particulier de cette classe est examiné en détail : le Trap2-Mixed. Il est démontré dans l'article qu'il est aussi précis et plus stable que les autres schémas, notamment en présence d'advection. Des tests numériques ont confirmé les analyses dans un contexte plus réaliste.

RK-IMEX HEVI schemes for fully-compressible atmospheric models with advection: analyses and numerical testing

Charles Colavolpe*, Fabrice Voitus and Pierre Bénard

CNRM UMR 3589, Météo-France/CNRS, Toulouse, France

*Correspondence to: Ch. Colavolpe, CNRM/GMAP, 42 av Gaspard Coriolis, F-31057 Toulouse Cedex, France. E-mail: charles.colavolpe@meteo.fr

The integration of the fully compressible non-hydrostatic equations with horizontally-implicit schemes creates scalability problems for massively parallel computing architectures. An alternative is to use horizontally explicit and vertically implicit (HEVI) approaches, where the implicit problems are only along the vertical direction. Besides, various multi-stage implicit-explicit (IMEX) methods, based on Runge-Kutta (RK) schemes, have been developed over recent years in order to achieve time-discretizations free of any computational modes, and possessing specified properties (accuracy-order, number of implicit iterations, amount of storage,...). This paper compares the analytical responses of three RK-IMEX HEVI schemes (identified as attractive for atmospheric modelling), for a linear fully compressible system supporting gravity and acoustic waves, as well as advection. Each scheme is analysed in two variants "UFPreF" and "UFPreB" recently proposed in the literature. The propagation of gravity waves is found to be generally well represented, but the advection makes unstable all UFPreB variants, which were on the contrary more stable without advection. The instability is analysed in a one-dimensional framework, and a new class of schemes is proposed to circumvent the problem (using four Butcher tableaux, at no extra cost). A particular member of this class is examined in detail: it is shown to be accurate and stable even with advection. Some numerical testing is provided to support the analyses in a more realistic context.

Key Words: HEVI scheme; RK-IMEX scheme; numerical analysis; advection; gravity waves;

Received . . .

1. Introduction

In numerical atmospheric modelling, considerable efforts are currently made to develop time discretization schemes able to fully exploit the performances of massively-parallel distributed-memory computers. In order to reach this requirement of scalability, such numerical algorithms should have minimal communications between the parallelized tasks, as well as efficient and balanced computations inside each task. The problem is complicated by the now generalized use of non-hydrostatic (NH) compressible systems in replacement of the Hydrostatic Primitive Equations (HPE) system, for high-resolution atmospheric applications. The dynamic of these NH compressible systems allows the propagation of vertically-propagating acoustic modes, which makes the governing equations more stiff than HPEs, and thereby more difficult to handle numerically.

In the case where the stiffness of a dynamical system only comes from a limited part of all the processes that are supported, Runge-Kutta Implicit-Explicit (RK-IMEX) schemes offer an attractive option (Pareschi and Russo 2001). RK-IMEX schemes are a wide class of time-discretizations which [in opposition to time-splitting methods, see e.g. Wicker *et al.* (2002)] use the same time-step for all terms or processes, and are inherently free of computational modes. Furthermore, RK-IMEX schemes use a multi-stage evaluation of intermediate states (termed "sub-stages" hereafter) that may be combined together to compute the state at the next time-step, which implies a storage of these intermediate stages. Moreover, they also allow a separate treatment for an explicitly-treated part and an implicitly-treated part of the system to be solved, in the computation of each intermediate state (to alleviate the above mentioned stiffness problem). A mathematical description of what is a multi-stage RK-IMEX scheme is provided in section 2.2. Although the absence of computational mode is guaranteed in these schemes, their relative complexity makes desirable a careful analysis of their properties (stability, accuracy, ...) prior to any practical use.

Some particular types of RK-IMEX schemes have been advocated and used for parabolic diffusion-convection systems (Ascher *et al.* 1997), recently prompting atmospheric modellers to also consider this class of schemes for the numerical integration of fully compressible non-hydrostatic systems. Commonly, for this new field of application, the implicitly-treated part of

© 2016 Royal Meteorological Society

the system is intended to contain at minimum the terms contributing to the vertical propagation of fast waves. This approach, termed "Horizontally Explicit and Vertically Implicit" or "HEVI" (Satoh 2002), is justified by the relatively large ratio of spatial resolution along horizontal and vertical directions encountered in atmospheric models: the explicit treatment of all horizontal processes makes it possible to use horizontally local (hence scalable) spatial discretizations, and the severe Courant-Friedrichs-Lewy (CFL) restriction on the time-step induced by vertically propagating waves is avoided by the implicit treatment along this direction. The overall stability of the resulting schemes is therefore expected to be mainly limited by the horizontal propagation of fast waves.

There is a growing interest for this kind of HEVI splitting approach combined with RK-IMEX time-stepping in the Numerical Weather Prediction (NWP) community. For instance, Ullrich and Jablonowski (2012) have examined three RK-IMEX HEVI schemes for non-hydrostatic compressible systems using a finite-volumes spatial discretization: a so-called "crude- splitting", the Strang-carryover splitting, and the so-called "ARS" scheme of Ascher *et al.* (1997). Besides, Giraldo *et al.* (2013) have studied the accuracy and efficiency of RK-IMEX HEVI methods spatially discretized by discontinuous Galerkin methods, for global and limited-area non-hydrostatic three-dimensional (3D) systems.

More recently, Weller *et al.* (2013) (WLW13 hereafter), and Lock *et al.* (2014) (LWW14 hereafter) have performed linear stability analyses of various RK-IMEX schemes proposed in earlier literature, for two alternative HEVI variants applied to a two-dimensional (2D) vertical-plane linear system supporting only acoustic waves and a trivial stationary mode:

- The "UFPreF" (U-Forward/Pressure-Forward) variant, where all horizontal source terms are solved explicitly and other terms acting on the vertical are solved implicitly.
- The "UFPreB" (U-Forward/Pressure-Backward) variant, where at a given sub-stage, the completed values of the horizontal wind-components are used to compute the pressure equation sources, thereby potentially enhancing the stability (see section 2.3).

From these analyses, three HEVI RK-IMEX schemes offering the best stability and accuracy were identified as the so-called "UJ3(1,3,2)", "ARK2(2,3,2)", "Trap2(2,3,2)(-1)" schemes. All three schemes have been specifically developed for use in atmospheric models. By contrast, schemes identified from the wider literature exhibited adverse properties: higher accuracy reached at an impractical cost, or inability to guarantee stability for all resolved wave-numbers within some restricted time-step. In addition, LWW14 pointed out that the "UFPreB" variant is more stable than the "UFPreF", for the simple 2D linear system examined therein.

However, as mentioned above, before considering these RK-IMEX HEVI methods as a viable fall-back solution for real NWP applications, further knowledge about their behaviour is needed, and this task must be undertaken for models capturing reasonably well the wide variety of processes acting in a real meteorological system.

The aim of this paper is (i) to extend LWW14 analyses to the case where gravity waves and advections are present in the system (for the best schemes identified therein); (ii) to demonstrate that gravity waves do not create serious problems but advection makes unstable all UFPreB schemes (they were more stable than UFPreF without advection); (iii) to propose a scheme which, although similar to Trap2(2,3,2)(-1), solves the instability linked to the presence of advection, and (iv) to assess the behaviour of all schemes (those selected from LWW14, and the proposed one) in a numerical model for which the framework is closer to real atmospheric conditions than in analyses.

This paper is organized as follows. The model equations and the framework of the numerical analyses are described in section 2. In section 3, the stability and the phase-error analysis of the aforementioned most promising RK-IMEX HEVI schemes are studied when applied to a system that supports both acoustic and gravity wave propagation but no advection. The response of the schemes is examined in section 4 for the same linear system but now including a basic-state horizontal advection. The instability of UFPreB schemes in presence of advection is investigated for a one-dimensional system in section 5. A class of new schemes is defined and one in particular is proposed and analysed in section 6. Numerical experiments supporting the validity of the analyses in a numerical model are presented in section 7. Finally, section 8 includes a brief summary and concluding remarks.

2. Framework for analyses

2.1. *Continuous system*

The governing system used in all subsequent analyses is the fully compressible Euler equations system in a 2D vertical Cartesian plane (x, z). For analyses, equations are linearized around a reference state $\overline{X}$ which is assumed to be isothermal (at temperature $\overline{T}$) and hydrostatically-balanced $d\overline{p}/dz = -\overline{\rho}\mathrm{g}$, where the pressure $\overline{p}(z)$ and density $\overline{\rho}(z)$ profiles are horizontally homogeneous and g is the acceleration due to gravity. The flow in $\overline{X}$ consists in a uniform horizontal wind $\overline{U}$. A linearisation of the fully compressible system around this basic-state yields

$$\partial_t u + \overline{U}\partial_x u + \partial_x P = 0, \qquad (1)$$

$$\partial_t w + \overline{U}\partial_x w - b + \left(\partial_z - \frac{\kappa}{\overline{H}}\right) P = 0, \qquad (2)$$

$$\partial_t b + \overline{U}\partial_x b + \overline{N}^2 w = 0, \qquad (3)$$

$$\partial_t P + \overline{U}\partial_x P + \overline{c}_s^2 \left[\partial_x u + \left(\partial_z - \frac{1-\kappa}{\overline{H}}\right) w\right] = 0, \qquad (4)$$

where (1)–(4) are respectively the two components of the linearised momentum equations for the perturbed velocity components u and w, the linearised thermodynamic equation expressed in term of the buoyancy variable $b = \mathrm{g}\theta/\overline{\theta}$, and the linearised continuity equation expressed in term of perturbed pressure potential variable $P = p/\overline{\rho}$. The reference potential temperature profile is defined by $\overline{\theta}(z) = \overline{T}(p_{00}/\overline{p})^{\kappa}$, where p_{00} is a constant, $\kappa = R/C_p$, R is the universal gas constant and C_p the specific heat of air at constant pressure. In (1)–(4), $\overline{H} = R\overline{T}/\mathrm{g}$ is the height-scale of the reference atmosphere, $\overline{c}_s^2 = \mathrm{g}\overline{H}/(1-\kappa)$ and $\overline{N}^2 = \kappa\mathrm{g}/\overline{H}$ are respectively the reference-state squared sound speed and buoyancy frequency.

This linear system retains the broad spectrum of time scales of the processes acting in the fully compressible Euler equations: fast processes are represented by acoustic and gravity waves propagation, and slow processes are represented by buoyancy and advection. The stiffness of the governing equations to be solved ultimately is therefore well captured by this linear system.

The system (1)–(4) may be written symbolically

$$\partial_t X = \overline{\mathcal{L}}.X, \qquad (5)$$

where the linear operator $\overline{\mathcal{L}}$ may be expressed as a symbolic matrix of operators

© 2016 Royal Meteorological Society

$$\overline{\mathcal{L}} = \begin{bmatrix} -\overline{U}\partial_x & 0 & 0 & -\partial_x \\ 0 & -\overline{U}\partial_x & 1 & -\left(\partial_z - \frac{\kappa}{\overline{H}}\right) \\ 0 & -\overline{N}^2 & -\overline{U}\partial_x & 0 \\ -\overline{c}_s^2\partial_x & -\overline{c}_s^2\left(\partial_z - \frac{1-\kappa}{\overline{H}}\right) & 0 & -\overline{U}\partial_x \end{bmatrix}, \tag{6}$$

and $X(x,z,t) = (u,w,b,P)$ is a symbolic state-vector of prognostic variables varying in time and space.

For an idealised vertically unbounded atmosphere, solutions of (1)–(4) that are oscillatory in time may be shown to necessarily satisfy

$$X(x,z,t) = \hat{X}_0 \exp\left\{i\left[k_x x + \left(k_z + i/2\overline{H}\right)z - \omega t\right]\right\}, \tag{7}$$

where $\hat{X}_0 \in \mathbb{C}^4$, $i^2 = -1$ and k_x, k_z, ω are three real numbers. This time-space structure characterizes an eigenmode for which k_x is the horizontal wave-number of the mode, k_z acts as a vertical wave-number, and ω is the frequency.

By inserting (7) in (1)–(4), and after some algebra the dispersion relation for this eigenmode writes

$$\hat{\omega}^4 - \overline{c}_s^2\left(k_x^2 + k_z^2 + \frac{1}{4\overline{H}^2}\right)\hat{\omega}^2 + \overline{c}_s^2\overline{N}^2 k_x^2 = 0, \tag{8}$$

where $\hat{\omega} = \omega - Uk_x$. This dispersion relation always admits four real roots, each pair of roots having opposite signs so that $\hat{\omega} \in \{\pm\hat{\omega}_a, \pm\hat{\omega}_g\}$, where $\hat{\omega}_a > \hat{\omega}_g > 0$. Although all waves of this system are indeed "mixed" waves, those associated to $\hat{\omega}_a$ and $\hat{\omega}_g$ will be referred to "acoustic" and "gravity" waves respectively.

Following the same approach and notation as in LWW14, for any eigenmode X satisfying (7) an "amplification" factor A_{tc} between two given times t and $t+\Delta t$ may be defined as $X(t+\Delta t) = A_{tc}X(t)$ (the subscript "tc" refers to the time-continuous system considered here). This complex number describes the phase shift as well as the amplification during the considered time interval. It takes the form $A_{tc} = e^{-i\omega\Delta t}$ with $\omega \in \{Uk_x \pm \hat{\omega}_a, Uk_x \pm \hat{\omega}_g\}$, which simply confirms that the modulus of the amplification factor is neutral $|A_{tc}| = 1$ for all four modes, if their spatial structure satisfies that given in (7). Their corresponding phase-shift $\omega\Delta t = -\mathrm{Arg}(A_{tc})$ (simply termed "phase" in LWW14, and hereafter as well, for consistency) may be expressed, after some algebra, by

$$\omega_a^{\pm}\Delta t = M_U\Delta t^* \pm \Delta t^*\left(1 + r^2 + \epsilon^2\right)^{1/2} \times \left\{\frac{1}{2} + \left[\frac{1}{4} - \frac{\overline{N}^2\Delta t^2}{\Delta t^{*2}\left(1 + r^2 + \epsilon^2\right)^2}\right]^{1/2}\right\}^{1/2}, \tag{9}$$

$$\omega_g^{\pm}\Delta t = M_U\Delta t^* \pm \Delta t^*\left(1 + r^2 + \epsilon^2\right)^{1/2} \times \left\{\frac{1}{2} - \left[\frac{1}{4} - \frac{\overline{N}^2\Delta t^2}{\Delta t^{*2}\left(1 + r^2 + \epsilon^2\right)^2}\right]^{1/2}\right\}^{1/2}, \tag{10}$$

where $M_U = \overline{U}/\overline{c}_s$ is the Mach-number of the basic flow, $r = k_z/k_x$ is the aspect ratio of the mode (horizontal scale over vertical scale), $\epsilon = 1/(2\overline{H}k_x)$ is an additional aspect ratio of the horizontal scale over the characteristic height-scale associated with the combined effect of buoyancy and Boussinesq terms, and $\Delta t^* = \overline{c}_s k_x \Delta t$ is a non-dimensional number, which, in view of time-discrete systems, would correspond to a horizontal wave-CFL number.

Following an argument in LWW14 [*], the ratio r is allowed to vary in the range $\left[10^{-2}, 10^3\right]$, and results are examined only for horizontal wave-CFL numbers Δt^* varying in the interval $[0,3]$, all schemes being unstable beyond this upper bound of Δt^*.

Fig. 1 represents the exact acoustic $\omega_a\Delta t$ and gravity $\omega_g\Delta t$ phase-shift magnitude with respect to r (abscissa) and Δt^* (ordinate), in absence of advection ($\overline{U} = 0$). Note that the very small phase-shift of gravity waves in the right panel have been multiplied by a factor of 32.

2.2. *RK-IMEX time-discretization*

The class of RK-IMEX time discretization schemes may be defined as follows. First a partitioning of the RHS terms of the system to be solved is introduced through

$$\partial_t X = \mathcal{E}(X) + \mathcal{I}(X), \tag{11}$$

where the term $\mathcal{E}$ denotes the part of the system RHS to be treated explicitly, and $\mathcal{I}$ the part of the system RHS to be treated implicitly (the exact content of these parts will be detailed in section 2.3). Then, two different multi-stage RK schemes are respectively applied to $\mathcal{E}$ and $\mathcal{I}$ parts. The RK scheme applied to $\mathcal{E}$ is purely explicit whereas that applied to $\mathcal{I}$ allows implicit evaluations at each sub-stage. The result may be written under the general form

$$\frac{X^{(j)} - X^0}{\Delta t} = \sum_{i=1}^{j-1}\widetilde{a}_{ji}\mathcal{E}\left[X^{(i)}\right] + \sum_{i=1}^{j}a_{ji}\mathcal{I}\left[X^{(i)}\right], \tag{12}$$

$$\frac{X^{+} - X^0}{\Delta t} = \sum_{j=1}^{\nu}\widetilde{b}_j\mathcal{E}\left[X^{(j)}\right] + \sum_{j=1}^{\nu}b_j\mathcal{I}\left[X^{(j)}\right], \tag{13}$$

where $\nu \geq 2$ is the total number of sub-stages of the RK-IMEX scheme, i,j are integer indices such as $1 \leq i \leq j \leq \nu$, $X^{(j)}$ denotes the value of the state variable at the j-th sub-stage, and the superscripts "0" and "+" correspond to the values of the state variable at times t and $t+\Delta t$ respectively. Notations like $\mathcal{E}\left[X^{(i)}\right]$, indicate that the terms of the sub-system $\mathcal{E}$ are evaluated using the state variable at sub-stage $X^{(i)}$,

The coefficients $\mathcal{A} = (a_{ji})$, $\widetilde{\mathcal{A}} = (\widetilde{a}_{ji})$ for $(i,j) \in [1,\nu]\times[1,\nu]$, and the weight-vectors $(b_j,\ c_j = \sum_{i=1}^{j}a_{ji})$ and $(\widetilde{b}_j,\ \widetilde{c}_j = \sum_{i=1}^{j}\widetilde{a}_{ji})$ for $j \in [1,\nu]$ may be classically represented by a double Butcher tableau:

$$\begin{array}{c|c} \widetilde{c} & \widetilde{\mathcal{A}} \\ \hline & {}^{\mathrm{T}}\widetilde{b} \end{array} \qquad \begin{array}{c|c} c & \mathcal{A} \\ \hline & {}^{\mathrm{T}}b \end{array},$$

where the left superscript ^{T} denotes the transpose operator. The first Butcher tableau defined by $(\widetilde{\mathcal{A}}, \widetilde{b}, \widetilde{c})$ describes the explicit part so that $\widetilde{a}_{ij} = 0$ for $i \geq j$, and the second one $(\mathcal{A}, b, c)$ corresponds to the implicit part of the scheme. RK-IMEX schemes, with their double Butcher tableau, are traditionally labelled in literature with the nomenclature [NAME] $k(s,\sigma,p)$, where k denotes the order of accuracy of the explicit part, s, the number of implicit inversions to be performed in the implicit part (i.e. the number of non-zero diagonal coefficients in $\mathcal{A}$), σ, the storage factor (i.e. the minimal number of explicit sub-stages that need to be stored to complete the time-step), and p, the overall order of accuracy of the scheme. The particular RK-IMEX schemes that will be

[*] The argument is based on the fact that in a discrete model, the resolved wavenumbers k_x, k_z are bounded by the physical meshes Δx, Δz. Moreover, the results shown below indicate *a posteriori* that all interesting features of the schemes are well-captured with this choice of interval for r (see left and right edges of Figs. 2, 3 and 7).

© 2016 Royal Meteorological Society

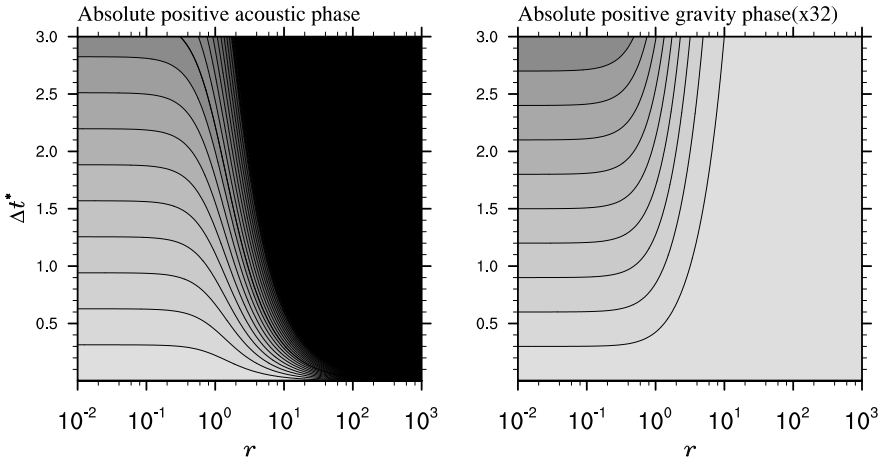


Figure 1. Exact phase-shift magnitude (shadings and contours) of fast waves described in the text as a function of r (abscissa) and Δt^* (ordinate), in the absence of mean flow ($\overline{U}=0$). Contour interval 0.1π. Left panel: phase-shift magnitude for acoustic waves (the black region on the right of the panel represents phase-shifts larger than 2π, very rapidly growing toward the upper-right corner, where the value would be $\approx 10^3\pi$). Right panel: phase-shift magnitude multiplied by 32 for gravity waves.

analysed in this paper, are the so-called UJ3(1,3,2), ARK2(2,3,2), and Trap2(2,3,2)(-1) scheme, according to this nomenclature. The Butcher tableaux for each of these three schemes are given in Appendix A.

The UJ3(1,3,2) RK-IMEX scheme, proposed by Ullrich and Jablonowski (2012), is based on a strong stability preserving RK3 third-order explicit scheme with only one implicit inversion (at the final stage). The ARK2(2,3,2) scheme has been designed by Giraldo *et al.* (2013) as a "L-stable second-order explicit time-stepping scheme. It has been shown to be the most stable RK-IMEX scheme from the preliminary numerical stability analysis of LWW14. However, ARK2(2,3,2) scheme is more expensive than UJ3(1,3,2) scheme since it requires one additional implicit inversion. Originally suggested by Wood *et al.* (2013) and slightly modified by LWW14, the Trap2(2,3,2)(-1) scheme is also a second-order explicit scheme, which is more stable than the UJ3(1,3,2) scheme and less expensive than the ARK2(2,3,2) scheme in term of storage factor.

2.3. *HEVI schemes for atmospheric modelling*

The various terms in the RHS of the system to be solved must now be partitioned between $\mathcal{E}$ and $\mathcal{I}$. Many options are possible, but atmospheric modelling usually retains the HEVI approach, which is defined by the only constraint that no implicit problem is to be solved spatially along the horizontal direction. For this, the partitioning must respect some basic conditions. The simplest option is obviously to impose that $\mathcal{I}$ can contain only terms with no horizontal dependency (they may have vertical dependency or not). This option is termed "UFPreF" in LWW14, and is the basic option which has been used by all earlier studies about HEVI schemes. However, LWW14 pointed out that it is possible to define options for which some terms are inserted in the $\mathcal{I}$ part, even if they possess a horizontal dependency. However, for such a scheme to remain an "HEVI" scheme (in the above sense) some constraints must be fulfilled. Simply speaking, this is possible only if the insertion of terms with horizontal dependency into the $\mathcal{I}$ part does not create closed coupling loops (as if e.g. $u^{(j)}$ requires the knowledge of $P^{(j)}$, and $P^{(j)}$ requires the knowledge of $u^{(j)}$, though the coupling loops could be more complicated, involving more variables than the two of this example). Indeed, the variant "HEVI UFPreB" proposed by LWW14 inserts only one term with horizontal dependency in the $\mathcal{I}$ part, thereby obviously fulfilling the above condition. An additional, similar variant, termed "HEVI UBPreF", has been examined in the present study.

For a non-linear atmospheric model, the partitioning between $\mathcal{E}$ and $\mathcal{I}$ parts would obey the following guidelines: The $\mathcal{E}$ part should at least contain all advection terms; the $\mathcal{I}$ part should at least contain the terms linked to the propagation of vertical acoustic waves, which includes the vertical parts of divergence and pressure gradient and also buoyancy and Boussinesq terms. Beside this, other terms would have an ambiguous status (as e.g. Coriolis, and orographic terms) and the response of the model may be dependent on the choices made here. The horizontal parts of divergence and pressure gradient also pertain to this set of terms with an ambiguous status, hence the three variants UFPreF, UFPreB and UBPreF mentioned above.

Consequently for the linear system (1)–(4) considered here, the content of $\mathcal{E}$ or $\mathcal{I}$ for HEVI schemes is defined by

$$\mathcal{E}=\begin{bmatrix} -\overline{U}\partial_x & 0 & 0 & -(1-\gamma_p)\partial_x \\ 0 & -\overline{U}\partial_x & 0 & 0 \\ 0 & 0 & -\overline{U}\partial_x & 0 \\ -(1-\gamma_u)\overline{c}_s^2\partial_x & 0 & 0 & -\overline{U}\partial_x \end{bmatrix} \tag{14}$$

$$\mathcal{I}=\begin{bmatrix} 0 & 0 & 0 & -\gamma_p\partial_x \\ 0 & 0 & 1 & -\left(\partial_z-\frac{\kappa}{\overline{H}}\right) \\ 0 & -\overline{N}^2 & 0 & 0 \\ -\gamma_u\overline{c}_s^2\partial_x & -\overline{c}_s^2\left(\partial_z-\frac{1-\kappa}{\overline{H}}\right) & 0 & 0 \end{bmatrix}, \tag{15}$$

where (γ_u,γ_p) denote markers, respectively set to $(0,0)$ for UFPreF, $(1,0)$ for UFPreB, and $(0,1)$ for UBPreF (note that in these definitions, the subscript of γ refers to the *term* of which the horizontal derivative is treated implicitly, and not to the *equation* which contains an additional implicit term). It is readily checked that the UFPreF scheme has no horizontal dependencies in $\mathcal{I}$ and that the two other options (UFPreB and UBPreF) have only one term with horizontal dependency inserted in $\mathcal{I}$, hence no coupling loops also. By analogy with Forward-Backward schemes introduced by Mesinger (1977) an increase of the stability domain at no significant computational overcost may be potentially expected with the UBPreF and UFPreB variants, compared to an UFPreF scheme, which justifies the study of their properties here. Note however that the UBPreF option will not be discussed in detail below, since results appeared to be always similar to that UFPreB option.

© 2016 Royal Meteorological Society

2.4. *General principle of stability analyses*

The stability of the above time-discrete schemes will be analysed following the same methodology as in LWW14. The analysis evaluates the stability and phase error of space-continuous modes which have the same structure in space as the time-continuous modes (7), i.e.

$$X(t) = \mathbf{X}(t)\exp[ik_x x + i(k_z + i/2\overline{H})z]. \tag{16}$$

where $\mathbf{X}(t)$ is a simple column vector of four time-dependent complex numbers. As a consequence of the assumed form (16), the symbolic matrices of operators $\mathcal{E}$ and $\mathcal{I}$ become simple complex matrices $\mathbf{E}$ and $\mathbf{I}$ for each mode under examination. These complex matrices $\mathbf{E}$ and $\mathbf{I}$ are formally identical to (14)–(15) except that ∂_x and ∂_z operator symbols are now respectively replaced by the complex numbers ik_x and $ik_z - (1/2\overline{H})$.

The amplification matrices of the j-th sub-stage for $j \in [1,\nu]$ and the final stage of the RK-IMEX scheme are defined respectively by $\mathbf{X}^{(j)} = \mathbf{A}^{(j)}.\mathbf{X}^0$, and $\mathbf{X}^+ = \mathbf{A}.\mathbf{X}^0$. By inserting these forms into (12) and (13), the amplification matrices become

$$\mathbf{A}^{(j)} = \mathbf{1} + \Delta t\sum_{i=1}^{j-1}\widetilde{a}_{ji}\mathbf{E}.\mathbf{A}^{(i)} + \Delta t\sum_{i=1}^{j}a_{ji}\mathbf{I}.\mathbf{A}^{(i)}, \tag{17}$$

$$\mathbf{A} = \mathbf{1} + \Delta t\sum_{j=1}^{\nu}\widetilde{b}_{j}\mathbf{E}.\mathbf{A}^{(j)} + \Delta t\sum_{j=1}^{\nu}b_{j}\mathbf{I}.\mathbf{A}^{(j)}, \tag{18}$$

where $\mathbf{A}^{(i)}$ and $\mathbf{A}$ are now 4×4 complex matrices, and $\mathbf{1}$ is the 4×4 identity matrix. Note that (17) is an implicit definition which requires the inversion of $\left(\mathbf{1} - a_{jj}\Delta t\mathbf{I}\right)$. The four complex eigenvalues of $\mathbf{A}$, associated with the pairs of acoustic and gravity numerical modes supported by the time-discrete linear system are denoted by λ_l for $l \in \{1,2,3,4\}$. For a given choice of Δt^*, M_U and r values, the amplification and the numerical phase of each numerical mode is given respectively by the modulus $\Gamma_l = |\lambda_l|$ and the argument $\varphi_l = \mathrm{Arg}\,(\lambda_l)$ of the associated eigenvalue of $\mathbf{A}$. The overall stability of the scheme is given by

$$\Gamma(\Delta t^*, r, M_U) = \max\left[\Gamma_l(\Delta t^*, r, M_U)\right]_{l\in\{1,2,3,4\}}. \tag{19}$$

In the following, quantities denoted $\lambda_a^{\pm}$, $\Gamma_a^{\pm}$, $\varphi_a^{\pm}$ refer to the acoustic wave pair, the subscripts "+" or "-" being associated with the wave travelling in the same direction as that identified in (9)–(10). Similar notations with subscript "g" refer to gravity waves.

3. Stability and phase-error analysis : case without advection

The starting point of the analyses is to numerically compute the four eigenvalues of the amplification matrix $\mathbf{A}$ for a sample of parameters $(\Delta t^*, r)$ possessing a sufficient resolution to capture all appropriate details (1000×200, here). In this section, the analysed responses (amplification and phase) of gravity and acoustic waves are presented separately, since the required quality of the propagation is not necessarily the same for the two types of waves. For instance, a severe distortion of acoustic waves propagation may be acceptable especially if this type of wave is damped by the scheme. In contrast, the propagation of meteorologically important gravity waves should remain as realistic as possible. However, for the sake of conciseness, the responses for the two waves inside a pair will not be distinguished in the results. As a matter of fact, it has been carefully checked that, with no exception for the schemes examined here, the eigenvalues inside a given pair were either both real, or complex conjugated. As a consequence, inside a given pair, amplifications may differ only when the eigenvalues are both real, and phases may fail to be opposed only when one is π and the other is 0. In the case where amplifications differ inside a pair, the most informative value will be indicated for each scheme (if unstable, the most unstable; if stable but damping, the most damped).

As outlined in LWW14, the four eigenvalues $\{\lambda_1,...,\lambda_4\}$ provided by a numerical software are sorted in a purely conventional order, and are not directly attributable to the ordered quadruplet $(\lambda_a^+, \lambda_a^-, \lambda_g^+, \lambda_g^-)$ of the four physical waves under examination. For any given value of r, in the limit of small Δt^* the attribution is easy because the four eigenvalues tend toward their time-continuous counterparts (located on the complex unit circle), and are thus directly identifiable. When Δt^* increases, the attribution remains easy as long as the four (continuous) trajectories $\lambda(\Delta t^*)$ of the eigenvalues in the complex plane remain differentiable, since continuity arguments are then sufficient to perform the attribution. A potential problem occurs when two trajectories simultaneously exhibit a breaking point while crossing together at the same time. This was found to only occur when two non-real conjugated trajectories simultaneously reach the real axis and suddenly break their direction to follow, more or less briefly, two opposed paths along the real axis (or the inverse scenario, starting from the real axis). Since for a given physical mode, a constant numerical phase in some region of the spectrum represents a packet of stationary waves, the distinction between, say, λ_a^+ and λ_a^- then looses any meaning in this region and becomes of little practical relevance. When this unphysical phenomenon occurs, both components of the group velocity vector completely vanish, and this rather undesired feature is specifically detailed in the text.

The values of the amplification factor (Γ_a, Γ_g) and the numerical phase (φ_a, φ_g) for acoustic and gravity pairs are displayed in Fig. 2 for the UFPreF variant, and in Fig. 3 for the UFPreB. The responses for the alternative UBPreF formulation are found to be exactly identical to those of the UFPreB variant and therefore are not displayed. Each figure shows the results for the three RK-IMEX schemes under examination, as indicated in the topmost legends. Inside each group of four panels corresponding to a given scheme, the upper and lower row respectively refers to acoustic and gravity modes; the left and right column respectively depicts the amplification and the phase (the phase for gravity modes is always multiplied by a factor of 32). In panels showing the amplification, the growth-rate in regions of instability is represented by a gray scale at values $(1, 1.25, 2)$, and the damping-rate in stable regions is depicted by black dashed contours $(0.99, 0.95, 0.5, 0.1)$.

The numerical phases may be compared to the exact phases in Fig. 1. The pattern of their variation according to Δt^* and r provides a qualitative indication on the phase velocity and on the magnitude of the group velocity components. For panels showing phases in Figs. 2 and 3, the regions of overall instability are masked by a white shading, and regions where the numerical phase exhibits (unphysical) uniform values are shaded in black.

• *UJ3(1,3,2) UFPreF*

For large values of r, gravity waves are found to participate to the instability of the scheme together with acoustic waves of the same geometry. There is a slight distortion of numerical phases in the region $(r \geq 1, \Delta t^* \approx 1)$ for gravity waves. This indicates a reversal of the horizontal group velocity. Moreover, gravity waves are slightly damped in this region.

For acoustic waves, there is a region where the numerical phase is identically equal to π. In this region (shaded in black in Panel b), the group velocity vector vanishes, and the scheme is

© 2016 Royal Meteorological Society

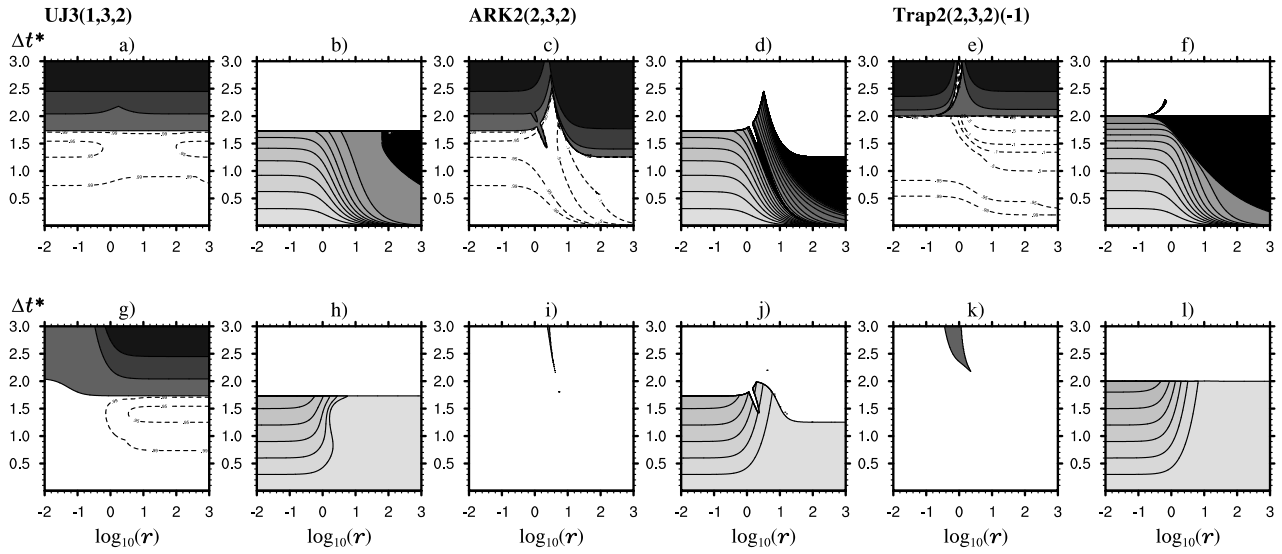


Figure 2. Amplification factor and numerical phase of acoustic waves (top panels) and gravity waves (bottom panels) for the three considered RK-IMEX UFPreF schemes indicated in the top legend. Abscissa : aspect ratio $r \in \left[10^{-2}, 10^{+3}\right]$, ordinates : horizontal wave-CFL number $\Delta t^* \in [0, 3]$. For each scheme, amplification factor (left panels) and positive numerical phase (right panels) are depicted. In panels showing amplification factors, regions between contours $\{1, 1.25, 2\}$ are shaded with increasingly dark gray colors; iso-contours $\{0.99, 0.95, 0.5, 0.1\}$ are indicated by black dashed-lines. For numerical phase panels, the contour interval is $\pi/10$ (phases for gravity waves are multiplied by a factor of 32); black areas depict regions where the numerical phase has a constant value, and white areas, regions of overall instability.

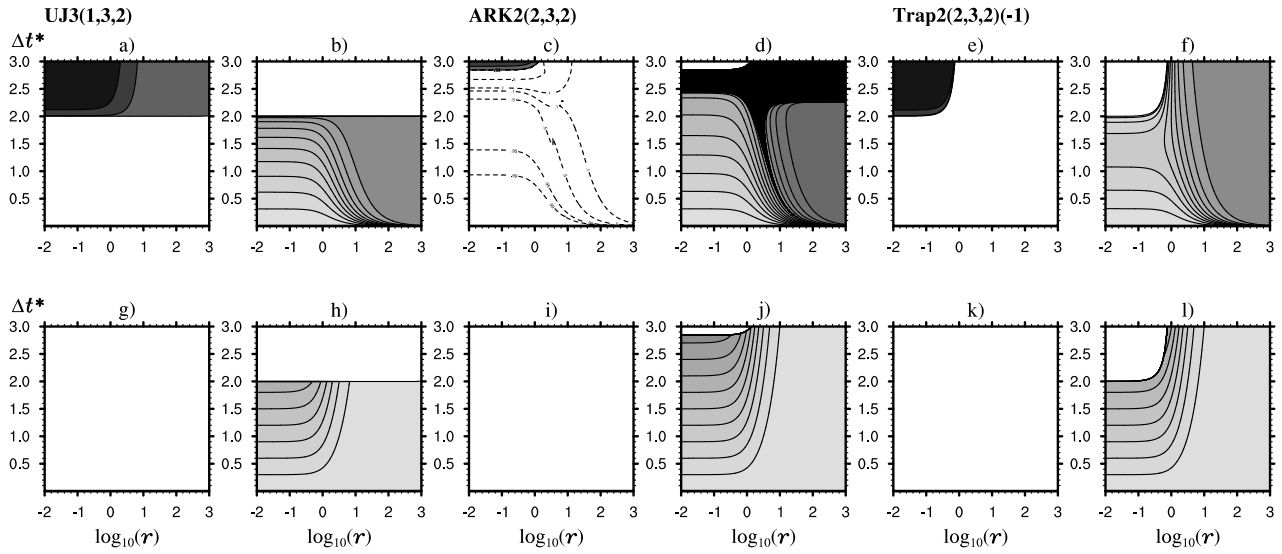


Figure 3. Same as Fig2, but for RK-IMEX UFPreB schemes.

likely to exhibit a noisy response, since the perturbations linked to some stationary source cannot be radiated away as they do in the real atmosphere or in a numerical scheme with appropriate dispersion properties. The occurrence of such a phenomenon does not necessarily mean that the scheme cannot be used in this region of the spectrum, but indicates that an increased level of damping or filtering may be needed to control the extra noise potentially occurring there.

• *UJ3(1,3,2) UFPreB*

The UFPreB variant brings two improvements : A correct behaviour for gravity waves is restored, and the region with a vanishing group velocity of acoustic waves disappears. This scheme provides appropriate responses, and a relatively large stability domain.

• *ARK2(2,3,2) UFPreF*

This scheme has a relatively smaller stability domain, originating from large values of the aspect ratio r, where $\Delta t^* < 1.25$ is required for stability of acoustic waves. Inside the stability domain, gravity waves are well behaved. Acoustic waves are damped almost everywhere, but especially for large values of r.

Similarly to UJ3(1,3,2) UFPreF, this scheme exhibits problematic (black shaded) regions where the numerical phase of acoustic waves is a constant, and where therefore the group velocity vector vanishes. In a first region (leftmost tongue-shaped black region in panel d), the phase is identically equal to π, and in a second region

© 2016 Royal Meteorological Society

(at the right of the panel) the uniform value is 2π. Although the width of the first region is limited the problem there deserves some care since for small values of r, the incriminated components approaching the stability limit $\Delta t^* \approx 1.47$ are not damped. For larger values of r (e.g. $r > 10$), these unphysical stationary waves are significantly damped, making the problem less critical.

• *ARK2(2,3,2) UFPreB*

The stability limit is the largest of all schemes examined here: $\Delta t^* \approx 2.8$. Moreover, for $\Delta t^* \lesssim 2.5$ both acoustic waves are damped, which may be viewed as an extra degree of safety. The propagation of gravity waves is not significantly distorted for $\Delta t^* \lesssim 2$, but for $\Delta t^* \gtrsim 2$, there is a slight delay. This delay becoming detectable is only an effect of the large stability domain, all other schemes would exhibit the same if their graphs were not masked by white shadings.

Here also, there is a (black-shaded) region where the group velocity vector vanishes for acoustic waves, their eigenvalues being identically real. Moreover, for $r \lesssim 1$, the components approaching the stability limit $\Delta t^* \approx 2.8$ are not damped. As a consequence, it is very likely that the viable domain of stability is indeed restricted to the region where all acoustic waves are damped, $\Delta t^* \lesssim 2.5$. Inside the tongue-shaped part of this black area, stationary waves are moderately to strongly damped ($\Gamma_a \lesssim 0.7 - 0.5$), and therefore less problematic.

• *Trap2(2,3,2)(-1) UFPreF*

The overall stability is relatively large ($\Delta t^* \lesssim 2$), and gravity waves are not significantly distorted. As in ARK2(2,3,2) UFPreF, the acoustic modes have a complex behaviour, combining damped responses and a large area of vanishing group velocity (indicated by black shadings).

For this scheme, the problem of vanishing group velocities is quite severe. The region of occurence is exactly identical to the one indicated by the highest curved contour in LWW14's Fig. 7.c (left panel). The severity of the problem comes from two main reasons. Firstly, the region where acoustic waves are stationary is large (in opposition to the limited size observed with other schemes); secondly, one of the two stationary waves in the acoustic pair is virtually not damped at all. This is not apparent in our figure because only the smallest damping response is plotted, but the two panels of LWW14's Fig. 6c clearly show this asymmetry. The stationary wave exhibits a damping factor larger than 0.95 at $\Delta t^* \approx 1.9, 1.3, 0.55$ for $r = 2, 10, 100$ respectively. There is therefore a risk that such a large region of the resolved spectrum with exactly stationary but neutral waves might require a very high (possibly unacceptable) level of artificial diffusion or filter to control the noise.

• *Trap2(2,3,2)(-1) UFPreB*

The overall stability ($\Delta t^* \lesssim 2$) is equal to the UFPreF variant, gravity waves propagation is not significantly distorted.

There is no region were the group velocity vector vanishes. However, the horizontal component of the group velocity vanishes for $r \lesssim 1$ and $\Delta t^* \approx 1.4$, while waves are not damped there.

Synthesis

In the absence of advection, the results of the linear analysis are not dramatically modified compared to LWW14, in spite of the addition of gravity waves in the linear system. The stability domains are unchanged, and gravity waves do not reduce the stability domain (they only participate to the instability in the UJ3(1,3,2) UFPreF sheme). The propagation of gravity waves is not adversely distorted except for UJ3(1,3,2) UFPreF again, for which damping and reversed horizontal group velocities are observed.

A salient result is that all UFPreF schemes and the ARK2(2,3,2) UFPreB scheme exhibit areas where the group velocity vector of acoustic waves completely vanishes, the eigenvalues lying on the real axis. The extension of the incriminated areas varies from one scheme to the other, and is the largest for the Trap2(2,3,2)(-1) UFPreF scheme. Most of the time, this undesirable feature is counterbalanced by a significant damping of the stationary waves, but for some of the schemes, the problem may also occur for almost neutral components. In any case, such adverse results, obtained in simple analyses are not a sufficient reason for rejecting a scheme without further examination of the behaviour in more realistic frameworks, but they may provide a guideline about the conditions in which problems are likely to occur, and they may suggest which type of experiment could be relevant to evaluate the practical impact of the identified problems in a real numerical model. We believe that this potentially adverse behaviour of most RK-IMEX HEVI schemes, though certainly present in LWW14 analyses was not identified and highlighted enough therein.

4. Stability analysis : case with horizontal advection

When a horizontal wind velocity u is present in a flow, classical results show that the phase velocities $\{-c, c\}$ of an acoustic or gravity wave pair in the resting fluid, become locally $\{u - c, u + c\}$ in the direction of the flow. Since horizontal propagation of fast waves is treated explicitly in HEVI schemes, a reasonable expectation for the stability limit Δt^* as a function of the local Mach number $M_u = u/c$ is therefore

$$\Delta t^*(M_u) = \Delta t^*(0)/(1 + M_u). \tag{20}$$

Smaller values would mean that the HEVI scheme fails to meet its initial objectives.

The modifications of analytical results brought by the addition of (uniform) advection terms in the explicit part $\mathcal{E}$ (see 11) are now examined, following the same methodology as in the previous section. In these analyses, only the overall stability as given by (19) is examined.

Figure 4 shows the maximum amplification factor (shadings) for each RK-IMEX HEVI scheme in UFPreF and UFPreB variants (top and bottom panels respectively), as a function $M_U \in [0, 1]$ (abscissa) and $\Delta t^* \in [0, 3]$ (ordinates), when r spans the interval $\left[10^{-2}, 10^3\right]$. Here also, results for the UBPreF variant, exactly identical to those obtained for UFPreB, are not shown. The stability is found to be always independent of the sign of M_U, hence only positive values of M_U are shown in the figures. The rather large upper limit $M_U = 1$ chosen here is suggested by the fact that for NWP models with a top domain boundary located at high levels in the atmosphere, stratospheric jets may occasionally approach the speed of sound (at least in analyses or forecasts), leading to Mach numbers actually close to unity. For bottom panels (UFPreB schemes), the limiting value $\Delta t^*(M_U)$ that should be expected according to (20) is indicated by a white dashed ark of hyperbola.

• *UFPreF schemes*

With UFPreF variant, the stability region for each of the three schemes is found to be in good agreement with what could be expected from (20): when the advective wind increases, the time-step has to be reduced accordingly. For the ARK scheme, the slightly less regular decrease with M_U is an indirect consequence of the irregular shape of the stability domain in Fig. 2c. The stability limit, which was significantly lower than for the two other

© 2016 Royal Meteorological Society

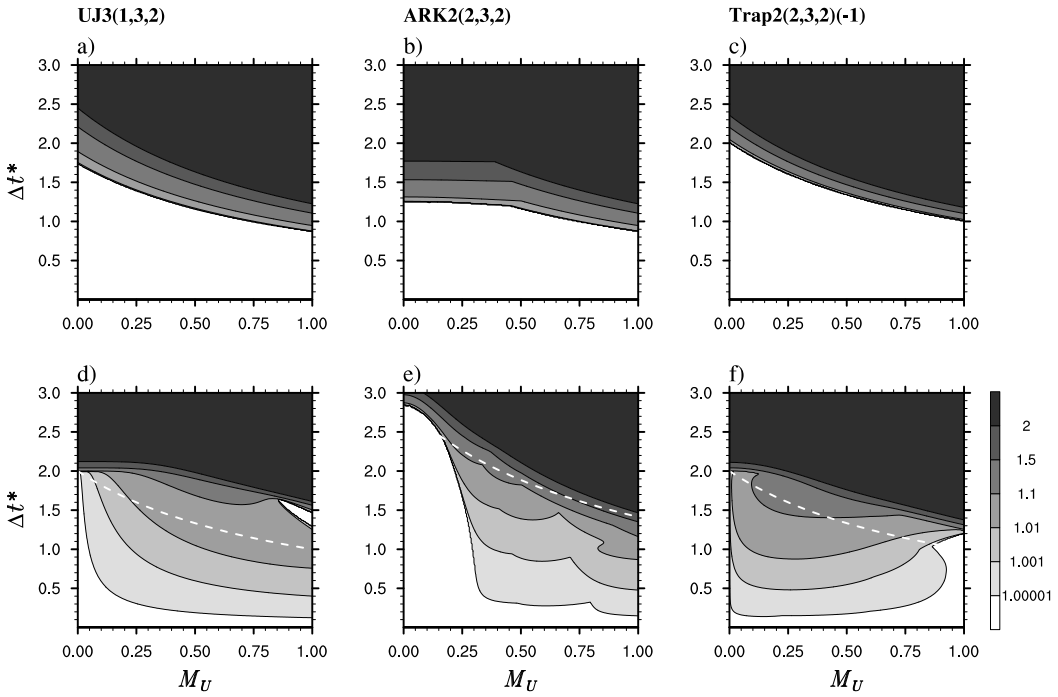


Figure 4. Maximum amplification factor for $r \in \left[10^{-2}, 10^{3}\right]$ wih respect to Δt^* and M_U for each schemes (UFPreF at top and UFPreB at bottom). The white dashed line in bottom panels depicts the expected decrease of the maximum time-step when M_U increases from zero, as given by (20).

schemes in the absence of advection, now becomes similar when M_U is likely to reach at least 0.4.

• *UFPreB schemes*

In UFPreB variant, all three schemes considered here are revealed to be more or less unstable, even at very small values of the Mach number for some of them. The poorest stability is observed with the Trap2(2,3,2)(-1) scheme: even for the unambitious values $\Delta t^* = 1$, the amplification factor is larger than 1.01 for almost any value of M_U larger than 0.1. Moreover, the regular and progressive growth of Γ along constant M_U lines in the bottom part of the graphs indicates that for a given time-step, the instability, not only restricted to the smallest scales, involves a large part of the resolved spectrum, which is a highly undesirable feature. †

For the UJ3(1,3,2) scheme, when $\Delta t^*(M_U)$ follows the expected upper limit (20) as indicated by the white dashed line, the amplification factor is about 1.01. This a better behaviour than for Trap2(2,3,2)(-1) scheme, but here also, an unstable amplification factor is unpleasantly observed for most of the components of the resolved horizontal spectrum.

For the ARK2(2,3,2) scheme, the use of a time-step approaching the expectation (white dashed line) leads to amplification rates of the order of 1.1 as soon as $M_U \gtrsim 0.3$. As a consequence, if the scheme is to be used in conditions where M_U is likely to exceed 0.3, the time-step will have to be chosen significantly smaller than that given by the reasonably expected limit (20). However, it must be outlined that if special care is taken to ensure that the wind is bounded to values obeying $M_U \lesssim 0.3$ in analyses and forecasts, the scheme might be viable for time-steps almost as large as those that can be expected from (20). The stability in this area of the spectrum is due to a significant damping of acoustic waves.

†For a given time-step, Δt^* simply becomes a mesure of the horizontal scale

Synthesis

For $|M_U| \leq 1$, the UJ3(1,3,2) and ARK2(2,3,2) schemes with UFPreF formulation remains stable provided that $\Delta t^* \lesssim 0.87$, whereas the Trap2(2,3,2)(-1) UFPreF scheme is stable for $\Delta t^* \lesssim 1$. As in the case with no advection the Trap2(2,3,2)(-1) scheme is the best scheme in term of stability range for the UFPreF variants. In UFPreB variants on the contrary, all RK-IMEX HEVI schemes examined here exhibit a significant level of instability even for relatively small Mach numbers. The ARK2(2,3,2) is the one with the widest region of stability for moderate Mach numbers $M_U \lesssim 0.3$, but a use in practical applications would then require to artificially bound fast stratospheric winds below this value. The growth rates encountered for nominal time steps (near the white dashed lines in Fig. 4), do not have large values and might therefore seem relatively harmless, but it must be taken into account that in modern NWP systems, forecasts have to be delivered up to a ten days range, which implies a huge number of time steps (the size of the time-step is significantly smaller than with semi-implicit schemes commonly used at present time). Furthermore, since the instability is not restricted to short wavelengths it might be very difficult to restore stability through damping mechanism without significantly affecting the flow itself, which would result in a detrimental effect on the quality of the forecast. Indeed, for all UFPreB schemes and most values of M_U the observed behaviour is that of unconditional instability.

The similarity between patterns in Fig.4 (bottom panels) suggests that the origin of this unstable behaviour is of the same nature for all UFPreB schemes considered herein. Moreover, a careful examination of the amplification rate as a function of r in the considered area (not shown) indicates that the instability reaches its maximum value (or almost) at $r = 0$, i.e. for very simple one-dimensional (1D) horizontal structures. In the next section, examination of a 1D horizontal sub-system extracted from (1)–(4), exhibiting the same unstable behaviour, but amenable to further algebraic analyses, is used to suggest possible solutions to the problem.

© 2016 Royal Meteorological Society

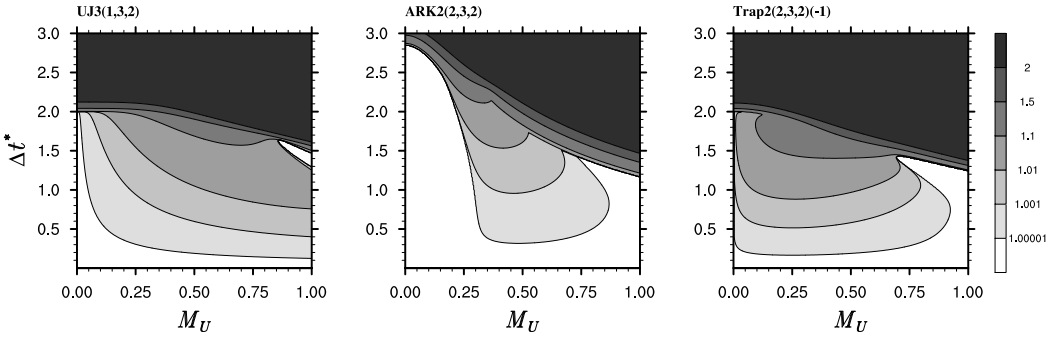


Figure 5. Amplification factor (contours and shadings) for analyses of the 1D acoustic system, as a function of M_U (abscissa) and Δt^* (ordinates). Times schemes are indicated in top legends, and UFpreB variant is used.

5. Stability of RK-IMEX schemes in 1D linear acoustic system with advection

The 1D linearized acoustic system with advection that will be analysed is simply an extract from (1)–(4)

$$\partial_t u + \overline{U}\partial_x u + \partial_x P = 0, \tag{21}$$
$$\partial_t P + \overline{U}\partial_x P + \overline{c}_s^2 \partial_x u = 0. \tag{22}$$

This time-continuous system admits time oscillating solutions of the form $(\hat{u}, \hat{P}) = (\hat{u}_0, \hat{P}_0)\exp\left\{i\left[k_x x - (\overline{U} \pm \overline{c}_s)k_x t\right]\right\}$, that corresponds to two waves propagating in the same and opposite directions as the advective wind ($\hat{u}_0$ and $\hat{P}_0$ are two complex numbers). As in section 2.4, the analyses consider the time-discrete evolution of modes $X(t) = \mathbf{X}(t)\exp(ik_x x)$ where $\mathbf{X}(t)$ is a simple column vector of two time-dependent complex numbers.

Since, by construction, there are no vertical dependencies, UFPreF variant of any RK-IMEX scheme here reduces to a fully explicit multi-stage treatment, i.e. all terms of (21)–(22) are dealt with the explicit Butcher tableau of the considered RK scheme. For the above modes $\mathbf{X}(t)$ the amplification matrix of any UFPreF (or explicit) RK scheme writes

$$\mathbf{A}^{\text{UFPreF}} = \mathbf{1} + ({}^{\mathrm{T}}\widetilde{b}\cdot e)\Delta t\mathbf{L} + ({}^{\mathrm{T}}\widetilde{b}\cdot\widetilde{c})\Delta t^2\mathbf{L}^2 + \widetilde{\Lambda}_1\Delta t^3\mathbf{L}^3 + \sum_{j=2}^{\nu}\widetilde{\Lambda}_j\Delta t^{j+2}\mathbf{L}^{j+2}, \tag{23}$$

where $\widetilde{\Lambda}_j = {}^{\mathrm{T}}\widetilde{b}\cdot\widetilde{\mathcal{A}}^j\cdot\widetilde{c}$ (for $j \geq 1$, $\widetilde{\mathcal{A}}^j$ being the j-th power of $\widetilde{\mathcal{A}}$), $e = (1,1)$, $\mathbf{1}$ is the 2×2 identity matrix, and

$$\mathbf{L} = -ik_x\overline{c}_s\begin{bmatrix} M_U & 1/\overline{c}_s \\ \overline{c}_s & M_U \end{bmatrix}. \tag{24}$$

Since the explicit part of all RK schemes considered herein is at least second order in time, ${}^{\mathrm{T}}\widetilde{b}\cdot e = 1$ and ${}^{\mathrm{T}}\widetilde{b}\cdot\widetilde{c} = 1/2$. Symmetrically, due to the fact that the explicit part of all RK schemes considered herein is at most third-order in time, it can be demonstrated that $\forall j \geq 2$, $\widetilde{\Lambda}_j = 0$, thus the last RHS term in (23) vanishes. Finally, $\widetilde{\Lambda}_1 = 1/6$ for UJ3(1,3,2) and ARK2(2,3,2) schemes, and $\widetilde{\Lambda}_1 = 1/4$ for Trap2(2,3,2)(-1) scheme.

Due to symmetries in the $\mathbf{L}$ and $\mathbf{A}^{\text{UFPreF}}$ matrices the two eigenvalues may be easily expressed in analytical form; then, after some algebraical manipulations, it appears that a necessary condition for the existence of a non-empty region of conditional stability in the spectrum is $\widetilde{\Lambda}_1 > 1/8$. For all RK schemes examined here, this condition is fulfilled by the coefficients of the explicit Butcher tableau. The upper limit of the horizontal wave-CFL number (Δt^*) for the explicit RK schemes can then be readily obtained

$$\Delta t^* \leq \frac{(2\widetilde{\Lambda}_1 - 1/4)^{1/2}}{\widetilde{\Lambda}_1\,(1+|M_U|)}. \tag{25}$$

As a result, the stability condition is given by $(1+|M_U|)\Delta t^* \leq \sqrt{3}$ for both UJ3(1,3,2) and ARK2(2,3,2) schemes, and by $(1+|M_U|)\Delta t^* \leq 2$ for Trap2(2,3,2)(-1) scheme. These analytical stability limits are in very good agreement with those obtained in the numerical analysis of the UFPreF schemes applied to the linear fully compressible system, except unsurprisingly for the ARK2(2,3,2) scheme, in which the stability limit had been found to be dependent on the aspect ratio r.

In UFPreB (or UBPreF) variants, the partitioning for the sub-system is the same as in section 2.3. Still following the principles of section 2.4, the resulting matrices $\mathbf{E}$ and $\mathbf{I}$ for a given mode k_x, k_z are therefore given by

$$\mathbf{E} = -ik_x\overline{c}_s\begin{bmatrix} M_U & (1-\gamma_p)/\overline{c}_s \\ (1-\gamma_u)\overline{c}_s & M_U \end{bmatrix}, \tag{26}$$
$$\mathbf{I} = -ik_x\overline{c}_s\begin{bmatrix} 0 & \gamma_p/\overline{c}_s \\ \gamma_u\overline{c}_s & 0 \end{bmatrix}. \tag{27}$$

The overall stability of the UFPreB (and UBPreF) schemes is evaluated numerically with the same approach as in section 4, the matrices $\mathbf{E}$ and $\mathbf{I}$ being now given by (26) and (27). Here also results for UBPreF variants are identical to UFPreB ones, and are not commented on further.

The amplification factor Γ (shadings) is displayed in Figure 5 as a function M_U (abscissa) and Δt^* (ordinates) for the three schemes indicated in top legends. The unstable behaviour observed in section 4, is very well captured in this simpler system, but now, the simplicity of the system allows further insight in the causes of the instability.

Since in (27) either γ_u or γ_p is zero for UFPreB and UBPreF schemes, an important property of the simple sub-system examined here is that $\mathbf{I}^2 = 0$, from which directly ensues $(\mathbf{1} - a_{jj}\Delta t\mathbf{I})^{-1} = (\mathbf{1} + a_{jj}\Delta t\mathbf{I})$. Consequently, the definition of the intermediate matrix $\mathbf{A}^{(j)}$ in (17) now becomes directly expressible as sums and products of known matrices. Making use of these simplifications, it appears after some algebra that the amplification matrices $\mathbf{A}$ of the three UFPreB schemes (denoted here by $\bullet$) may be written under the general form

$$\mathbf{A}_\bullet^{\text{UFPreB}} = \mathbf{A}_\bullet^{\text{UFPreF}} + \delta\mathbf{A}_\bullet, \tag{28}$$

where the deviations $\delta\mathbf{A}_\bullet$ now have simple expression in terms of sums and products

© 2016 Royal Meteorological Society

$$\delta\mathbf{A}_{\text{Trap2}} = \tfrac{\Delta t^4}{8}\left(\mathbf{IE}+\mathbf{EI}\right)\left(\mathbf{E}+\mathbf{I}\right)^2 + \tfrac{\Delta t^5}{16}\mathbf{IEIE}\left(\mathbf{E}+\mathbf{I}\right), \quad (29)$$

$$\delta\mathbf{A}_{\text{UJ3}} = \tfrac{\Delta t^3}{12}\left(\mathbf{E}^2\mathbf{I}+\mathbf{IE}^2+\mathbf{IEI}-2\mathbf{EIE}\right) + \tfrac{\Delta t^4}{12}\left(\mathbf{E}^3\mathbf{I}+\mathbf{IE}^3+\tfrac{3}{2}\mathbf{IE}^2\mathbf{I}\right) + \tfrac{\Delta t^5}{24}\mathbf{IE}^3\mathbf{I}, \quad (30)$$

$$\delta\mathbf{A}_{\text{ARK2}} = \left(3\sqrt{2}-4\right)\tfrac{\Delta t^3}{6}\mathbf{EIE} + \left(1-\tfrac{1}{\sqrt{2}}\right)\tfrac{\Delta t^4}{6}\left(\mathbf{E}+\mathbf{I}\right)\left(\mathbf{IE}+\mathbf{EI}\right)\left(\mathbf{E}+\mathbf{I}\right) + \left(1-\tfrac{1}{\sqrt{2}}\right)^2\tfrac{\Delta t^5}{6}\mathbf{EIEIE}. \quad (31)$$

The deviation of eigenvalues from an (unstable) UFPrefB scheme to its (stable) UFPreF counterpart clearly comes from the deviation matrix $\delta\mathbf{A}_\bullet$. The advantage of the above expressions is that the origins of any individual term or factor may be directly tracked up to the Butcher tableaux of the scheme.

The impact of each term in $\delta\mathbf{A}_\bullet$ on the stability may be accessed by performing the eigenvalues analysis with this term specifically set to zero in the expression of the amplification matrix. Applying this heuristic method reveals that the instability for small M_U values observed in UJ3(1,3,2) and Trap2(2,3,2)(-1) schemes is mostly linked to the term in Δt^5 in $\mathbf{A}_{\text{UJ3}}$ and $\mathbf{A}_{\text{Trap2}}$: cancelling this term restores the stability at small Mach numbers for both schemes, in a similar way as in the middle panel of Fig. 5. For the ARK2(2,3,2) scheme the term in Δt^5 is also present but the scheme is stable for small M_U. This suggests that some difference in the form of this term is responsible for the instability or stability. It is hypothesised here that the difference in stability is due to the nature of the matrix in the Δt^5 term. Another analysis of UJ3(1,3,2) and Trap2(2,3,2)(-1) schemes with the last two lines of the implicit Butcher tableau replaced by that of the explicit one led again to very similar results (as in the middle panel of Fig. 5), suggesting that a better combination of backward and forward treatments for u and P may solve the problem.

However, this kind of *ad hoc* manipulation, though possibly leading to enhanced stability, does not guarantee that the gain in stability is optimal or that the resulting scheme will be second-order in time. Indeed for UJ3(1,3,2) and Trap2(2,3,2)(-1), the scheme modified in this way is only first order in time.

The analyses of this section show that the stability of the linear 1D acoustic system with advection dramatically depends on the treatment of the adjustment $\partial_x u$ and $\partial_x P$ terms. In UFPreB or UBPreF variants, one of this two terms is always treated with the implicit Butcher tableau of the acoustic waves at each sub-stage, while the other term is always treated with the explicit Butcher tableau at each sub-stage. The above simple *ad hoc* manipulations indicate that this limitation may be far from an optimum as regards the stability, and that combining backward and forward treatments for $\partial_x u$ and $\partial_x P$ terms is likely to improve the stability. Now returning to 2D or 3D systems, the results of this section suggest the introduction of additional Butcher tableaux specifically dedicated to the terms that had been labelled as "ambiguous" when defining the partition between implicit and explicit Butcher tableaux in section 2.3. This possibility is examined in the next section.

6. RK-IMEX HEVI schemes with four Butcher tableaux: definition and an application

Drawing from the argument in previous section, a class of RK-IMEX HEVI schemes is proposed, in which terms involving the horizontal divergence and the horizontal pressure gradient may be subjected to their own backward/forward treatment, independently of the explicit and implicit treatments of slow and fast terms. Consequently, the scheme must have four Butcher tableaux and the system to be solved must be partitioned into four sub-systems (instead of two)

$$\partial_t X = \mathcal{E}'(X) + \mathcal{I}'(X) + \mathcal{U}(X) + \mathcal{P}(X). \quad (32)$$

The sub-system $\mathcal{E}'$ contains advection terms and all contributions that have to be treated exclusively in the explicit part of the scheme. The part $\mathcal{I}'$ contains the terms contributing to the vertical propagation of the acoustic modes and to be treated in the implicit part of the scheme. $\mathcal{U}$ and $\mathcal{P}$ contain the non-advective terms involving the horizontal divergence and the horizontal pressure gradient respectively. Concretely, for the system (1)–(4) the symbolic matrix $\mathcal{E}'$ contains only the diagonal part of $\mathcal{E}$ in (14), $\mathcal{I}'$ is equal to the UFPref version of $\mathcal{I}$ in (15), i.e. with γ_u and γ_p set to zero. Symbolic matrices $\mathcal{U}$ and $\mathcal{P}$ are identically zero except corner elements $\mathcal{U}_{41} = -\bar{c}_s^2\partial_x$ and $\mathcal{P}_{14} = -\partial_x$ respectively.

Then, the system partitioned as in (32) is time-discretized with a ν-stages RK-IMEX HEVI scheme where each sub-system $\mathcal{E}'$, $\mathcal{I}'$, $\mathcal{U}$ and $\mathcal{P}$ is treated with its own Butcher tableau, $\{\mathcal{A}^{\text{e}}, b^{\text{e}}, c^{\text{e}}\}$, $\{\mathcal{A}^{\text{i}}, b^{\text{i}}, c^{\text{i}}\}$, $\{\mathcal{A}^{\text{u}}, b^{\text{u}}, c^{\text{u}}\}$, and $\{\mathcal{A}^{\text{p}}, b^{\text{p}}, c^{\text{p}}\}$ respectively. The Butcher tableau $\{\mathcal{A}^{\text{e}}, b^{\text{e}}, c^{\text{e}}\}$ describes the explicit part of the RK-IMEX scheme, hence its diagonal elements are set to zero (i.e, $a^{\text{e}}_{jj} = 0$ for $j \in [1,\nu]$). The Butcher tableaux $\{\mathcal{A}^{\text{k}}, b^{\text{k}}, c^{\text{k}}\}$, $\text{k} \in \{\text{i},\text{u},\text{p}\}$ all involve implicit or backward evaluations, and therefore may have non-zero diagonal elements. However for the scheme to be actually an HEVI scheme, terms in $\mathcal{U}$ and $\mathcal{P}$ cannot be treated backward at the same sub-stage, as a result the diagonal element of the matrices $\mathcal{A}^{\text{u}}$ and $\mathcal{A}^{\text{p}}$ satisfy $a^{\text{u}}_{jj}a^{\text{p}}_{jj} = 0$, for $j \in [1,\nu]$. It is also assumed that the weight-vectors of all considered Butcher tableaux satisfy $c^{\text{k}}_j = \sum_{i=1}^{j} a^{\text{k}}_{ji}$ for $j \in [1,\nu]$; and $\sum_{j=1}^{\nu} b^{\text{k}}_j = 1$ for $\text{k} \in \{\text{e},\text{i},\text{u},\text{p}\}$. Finally, it can be shown that second-order accuracy in time (usually considered as desirable for NWP applications) is achieved provided that the weight-vectors of Butcher tableaux fulfill the condition ${}^{\text{T}}b^{\text{k}} \cdot c^{\text{k'}} = 1/2$, for $(\text{k},\text{k'}) \in \{\text{e},\text{i},\text{u},\text{p}\} \times \{\text{e},\text{i},\text{u},\text{p}\}$, (see Appendix B for a brief demonstration).

A particular member of the class defined above is now proposed, based on the Trap2(2,3,2)(-1) scheme. The resulting scheme is termed "mixed-Trap2(2,3,2)(-1)" hereafter. The main reason for choosing Trap2(2,3,2)(-1) as a starting point for this new scheme, is the efficiency of the algorithm and the fact that it may be viewed as a conceptually simple predictor-corrector algorithm.

The explicit and implicit Butcher tableaux of the new scheme are taken exactly identical to those of the original Trap2(2,3,2)(-1) scheme. The principle adopted here to build the two additional Butcher tableaux is to take alternatively one line from the explicit Butcher tableau and the next one from the implicit Butcher tableau (the detailed definition of the starting point is of no importance since switching the two tableaux leads to the same results). The final stage of the scheme is taken equal to the last sub-stage for all tableaux. The proposed scheme therefore has the following Butcher tableaux

(e)
$$\begin{array}{c|ccccc} 0 & 0 & & & & \\ 1 & 1 & 0 & & & \\ 1 & \frac{1}{2} & \frac{1}{2} & 0 & & \\ 1 & \frac{1}{2} & 0 & \frac{1}{2} & 0 & \\ \hline & \frac{1}{2} & 0 & \frac{1}{2} & 0 \end{array}$$

(i)
$$\begin{array}{c|ccccc} 0 & 0 & & & & \\ 1 & 1 & 0 & & & \\ 1 & \frac{1}{2} & 0 & \frac{1}{2} & & \\ 1 & \frac{1}{2} & 0 & 0 & \frac{1}{2} & \\ \hline & \frac{1}{2} & 0 & 0 & \frac{1}{2} \end{array}$$

(u)
$$\begin{array}{c|ccccc} 0 & 0 & & & & \\ 1 & 1 & 0 & & & \\ 1 & \frac{1}{2} & \frac{1}{2} & 0 & & \\ 1 & \frac{1}{2} & 0 & 0 & \frac{1}{2} & \\ \hline & \frac{1}{2} & 0 & 0 & \frac{1}{2} \end{array}$$

(p)
$$\begin{array}{c|ccccc} 0 & 0 & & & & \\ 1 & 1 & 0 & & & \\ 1 & \frac{1}{2} & 0 & \frac{1}{2} & & \\ 1 & \frac{1}{2} & 0 & \frac{1}{2} & 0 & \\ \hline & \frac{1}{2} & 0 & \frac{1}{2} & 0 \end{array}$$

© 2016 Royal Meteorological Society

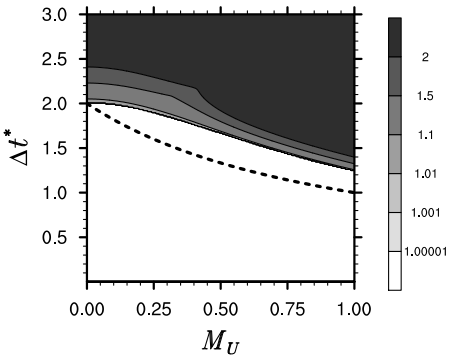


Figure 6. Same as Fig. 4 but for new mixed-Trap2(2,3,2) RK-IMEX HEVI alternative treatment. The dashed line indicates the stability limit of the Trap2(2,3,2)(-1) in UFPreF configuration.

The number of independent explicit sub-stages and the number of implicit sub-stages remain unchanged with respect to the original Trap2(2,3,2)(-1) scheme. Besides, the weight-vectors of the four Butcher tableaux RK-IMEX HEVI scheme fulfill the second-order overall accuracy condition (B.9) for the case of four Butcher tableaux. The scheme proposed here presents exactly the same characteristics as Trap2(2,3,2)(-1) scheme in term of storage factor and total number of implicit inversions. The mixed-Trap2(2,3,2)(-1) scheme is therefore as accurate as the original Trap2(2,3,2)(-1) at no overcost per time-step, and is shown below to be more stable

The overall stability of the mixed-Trap2(2,3,2)(-1) applied to the linear fully compressible system in presence of a linear advection (1)–(4) is assessed through the same numerical stability analysis as in section 4. Results of this analysis are displayed in Figure 6 in the same format as Fig. 4. The new scheme exhibits a wider domain of stability than all schemes in UFPreF or UFPreB variants examined in this paper. For $|M_U| \leq 1$, the maximum allowable horizontal wave-CFL number Δt^* is close to 1.25.

The stability and the phase-shift error for acoustic and gravity waves in absence of advection are also presented, as in section 3. Figure 7 depicts the result of this analysis in the same format as in Fig. 2 . The overall stability is almost completely controlled by the acoustic modes, and the maximum horizontal wave CFL for stability is $\Delta t^* \lesssim 2$. Inside the domain of stability the scheme is neutral for gravity modes, and their phases are well captured over all the range of r. As for the UFPreB Trap2(2,3,2)(-1) scheme, there is a region of the stable domain where the phase of the acoustic modes which remain constant. This region is delimited by $\Delta t^* > 1.7$ and $r \lesssim 1$, however, in a very large part of this region all acoustic modes are damped thereby alleviating the problem, except very close to the stability limit (at $\Delta t^* = 1.97$ the damping factor is about 0.95). There is also a very small area where the horizontal phase velocity vanishes (near $r < 10^0$ and $\Delta t^* \approx 1.6$), but here again the problem is restricted to values of Δt^* very close to the stability limit in the resolved spectrum.

These linear numerical analyses show that a clear benefit may be obtained through a specific treatment of horizontal adjustment terms, using additional implicit Butcher tableaux. The proposed mixed-Trap2(2,3,2)(-1) scheme appears to be a more attractive RK-IMEX HEVI time-stepping scheme than all those examined in LWW14. Returning to the view of Trap2(2,3,2)(-1) scheme as essentially a predictor-corrector algorithm, it may be hypothesized that the fact to apply the backward treatment alternatively to $\partial_x u$ and to $\partial_x P$ terms results in a more balanced progression of these two terms at successive sub-stages of the scheme than with the UFPreB variant.

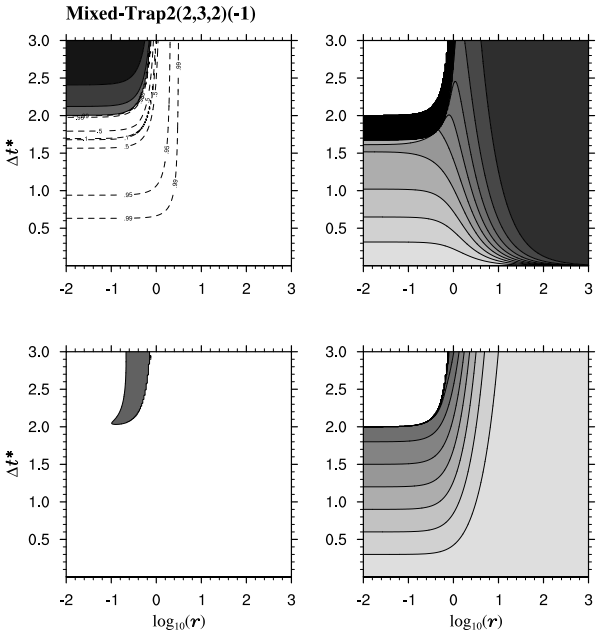


Figure 7. Same as Fig. 2 and 3, but for the proposed mixed-Trap2(2,3,2)(-1) scheme.

7. Assessment in numerical simulations of idealised flows

In order to assess the validity of the above analyses, a vertical plane $(x - z)$ numerical model has been developed, and simulations of the two idealised test cases proposed by Skamarock and Klemp (1994) (SK94, hereafter) have been performed: the non-hydrostatic gravity wave (NHGW) and hydrostatic inertia-gravity wave (HIGW). These two idealised test-cases give access to the behaviour of time-discrete schemes in a more realistic context than in analyses, while exploring two different aspect ratios of grid spacing and various Mach numbers for the ambient flow.

As in the linear system (1)–(4), the model still predicts the evolution of perturbations (u, v, w, b, P) around of a time-independent and horizontally-homogeneous atmospheric basic-state $(\overline{U}, \overline{T}, \overline{p})$. However the absolute temperature may now have a vertical variation in the basic-state $\overline{T} = \overline{T}(z)$. As a consequence some constant parameters of the linear system now become vertical profiles obeying

$$\overline{H}(z) = R\overline{T}/g, \tag{33}$$

$$d \ln \overline{p}(z)/dz = -1/\overline{H}, \tag{34}$$

$$\overline{\rho}(z) = \overline{p}/\left(R\overline{T}\right), \tag{35}$$

$$\overline{c_s}^2(z) = R\overline{T}/(1-\kappa), \tag{36}$$

$$\overline{N}^2(z) = g\left(d \ln \overline{\theta}/dz\right). \tag{37}$$

The governing equations of the 2D vertical plane numerical model are

$$D_t\, u + \partial_x\, P - f v = 0, \tag{38}$$

$$D_t\, v + f u = 0, \tag{39}$$

$$D_t\, w - b + \left[\partial_z + \overline{\Gamma}_w(z)\right] P = 0, \tag{40}$$

$$D_t\, b + \overline{N}^2(z)\, w = 0, \tag{41}$$

$$D_t\, P + \overline{c}_s^2(z) \left[\partial_x\, u + \left(\partial_z - \frac{1-\kappa}{\overline{H}(z)}\right) w\right] = 0, \tag{42}$$

© 2016 Royal Meteorological Society

where $D_t = \partial_t + (\overline{U}+u)\partial_x + w\partial_z$ and

$$\overline{\Gamma}_w(z) = -(d\ln\overline{T}/dz) - \kappa/\overline{H}.$$

The Coriolis parameter is set to $f = 10^{-4}\,s^{-1}$ for HIGW case and to zero for the NHGW case. The basic-state is geostrophically balanced with the tunable uniform flow $\overline{U}$, i.e. the prognostic equation for v only contains the perturbation of Coriolis force (fu). The surface temperature is taken as $\overline{T}(0) =$ 300K and the thermal profile is chosen such as the Brunt-Väisälä frequency is uniform $\overline{N} = 0.01\,s^{-1}$, i.e. $\overline{T}(z) = \overline{T}(0)[(1 - N_*^2/\overline{N}^2)\exp(\overline{N}^2 z/g) + N_*^2/\overline{N}^2]$, where $N_*^2 = g^2/\left[C_p\overline{T}(0)\right]$.

The test cases consist in simulating the evolution of a potential temperature perturbation

$$\theta'(x,z) = \Delta\theta_0 \frac{\sin(\pi z/z_T)}{1+(x-x_c)^2/a^2},$$

where $\Delta\theta_0 = 0.01\,\mathrm{K}$, $z_T = 10\,\mathrm{km}$ is the depth of the domain, $x_c = -60\,\mathrm{km}$ denotes the position of the perturbation centre. The domain is a channel with periodic boundary conditions along x, and rigid lower and upper boundary conditions at $z = 0$ and $z = z_T$.

For the NHGW case, the half-width of the perturbation a and the domain width L_x are set to $(a = 2500\,\mathrm{m}, L_x = 300\,\mathrm{km})$, and the mesh increments are $\Delta x = \Delta z = 500\,\mathrm{m}$. For the HIGW case, $a = 100\,\mathrm{km}$, $L_x = 5120\,\mathrm{km}$, $\Delta x = 10\,\mathrm{km}$ and $\Delta z = 100\,\mathrm{m}$.

All time-schemes examined in this paper have been implemented in the numerical model, moreover, simulations of the two cases with the standard fourth-order RK4 explicit scheme (using a small time-step) have also been performed as a reference. Coriolis terms are always treated in the explicit sub-system ($\mathcal{E}$ or $\mathcal{E}'$), all other terms are partitioned between relevant sub-systems as indicated above for each scheme (the partitioning for the v equation follows that of u equation).

As regards the model space-discretization, a Fourier spectral transform method is employed to compute horizontal derivatives and the collocation grid has the same number of degrees of freedom as the spectral space (no filtering). Along the vertical, a second-order finite-difference scheme is used with a staggered Charney-Phillips grid. Appendix C gives details of the variables placement, spatial operators, and the resulting 1D implicit vertical problem to be solved at relevant sub-stages. In all experiments performed herein, no numerical diffusion or artificial damping of any kind has been employed. For a given value of the time-step Δt and a given mesh-size Δx, the non-dimensional parameter Δt_* defined in analyses reaches the value

$$\Delta t^*_{\max} = \pi \Delta t\, \overline{c}_s(0)\,/\Delta x, \tag{43}$$

where $\overline{c}_s(0)$ is the maximum value of the sound speed (at ground).

During the simulations, the initial perturbation radiates waves symmetrically along x axis, and all disturbances travel with their own velocity relatively to the prescribed mean flow $\overline{U}$. The simulations are first performed with the same mean flow as in SK94 and Melvin *et al.* (2010) i.e $\overline{U} = 20\,\mathrm{m.s}^{-1}$ ($M_U \approx 0.06$). The time-step is chosen in such a way that $\Delta t^*_{\max} = 1$, and the length of integration is taken long enough for allowing possible instabilities to develop. As predicted by analyses, for these settings, all schemes provide stable integrations for NHGW and HIGW, except the Trap2(2,3,2)(-1) UFPreB and UBPreF schemes, in agreement with Fig. 4 (at $M_U \approx 0.06$, $\Delta t^* = 1$). All results obtained with stable schemes at 3000 s for NHGW case and 60000 s for HIGW case are qualitatively similar to that of the reference simulation and to previous simulations of the same cases using fully compressible systems with Boussinesq terms retained [e.g, Giraldo and Restelli (2008); Melvin *et al.* (2010)]. As an illustration, the numerical solution of the perturbed potential temperature for the proposed mixed-Trap2(2,3,2)(-1), is depicted in Fig 8.

The long-term stability of the schemes is then assessed through simulations of NHGW and HIGW cases up to 48 h (at this time, the evolution of the flow is still deterministic). For a given configuration (NHGW or HIGW) the time-step is chosen so as $\Delta t^*_{\max} \in \{1., 1.25, 1.5, 1.75, 2.\}$, and for each value of the time-step, the maximum value of the basic-state wind $\overline{U}$ for which the model remains stable after 48 hours is sought empirically. This maximum value is then converted into a basic-state Mach number through $M_U = \overline{U}/\overline{c}_s(0)$ for comparison with analytical results. It has also been checked that stable integrations produced results comparable to the RK4 simulations of reference.

The outcomes of this experimental study are given in Table 1. For NHGW case the maximum Mach number allowed by each scheme is in good agreement with that given by theoretical stability analyses. The Trap2(2,3,2)(-1) UFPreB scheme is found to be unstable for all experiments, even at $M_U = 0.01$. On the contrary, for some of the lowest values of $\Delta t^*_{\max}$, UJ3(1,3,2) and ARK2(2,3,2) UFPreB schemes lead to stable integrations whereas the growth-rate predicted by analyses is between 1.001 and 1.01. This slight discrepancy may originate from differences in the context as e.g. the bounded and non-isothermal atmosphere of the numerical model or the interpolations used in the vertical discretization. This is not investigated further.

The results for the HIGW case (bottom part of Table 1) are found very similar to the upper part. These results agree with those presented in WLW13 (at same resolutions) showing that Trap2(2,3,2)(-1) UFPreB is much less stable than other HEVI schemes in presence of advection. However results presented in WLW13 are restricted to very small values of M_U (around 0.04) and thereby, cannot make apparent the limitations predicted in Fig. 4 and confirmed in the Table, for UFPreB schemes in presence of large advection velocities. Since large advection velocities are commonly encountered in the real atmosphere, tests with low values of M_U such as those presented in WLW13 are arguably not demanding enough for validations in view of NWP.

Finally, the main result in Table 1 is that the stability of the Trap2(2,3,2)(-1) UFPreF scheme, which had been lost with the UFPreB scheme, is more than restored with the Mixed scheme. For $\Delta t^*_{\max} \geq 1.25$, the stability limit for M_U is in very good agreement with Fig. 6. For $\Delta t^*_{\max} = 1$, the maximum Mach number $M_U = 1.4$ is beyond the right edge of the figure.

8. Summary and concluding remarks

The response of some RK-IMEX HEVI schemes proposed in the literature for atmospheric modelling has been examined in a 2D governing system allowing gravity and acoustic waves, and advection. This is therefore an extension of LWW14 analyses, using a similar methodology. The schemes examined here were those identified as the most promising therein: UJ3(1,3,2), ARK2(2,3,2) and Trap2(2,3,2)(-1).

Without advection, the behaviour of the schemes considered was not substantially modified compared to LWW14, however the propagation of acoustic waves appeared to be corrupted (with vanishing group velocities in large areas of the resolved spectrum) for most of the schemes. This undesired feature was certainly present in LWW14 analyses, but was neither immediately identifiable in figures therein, nor mentioned in the text. It is conjectured that when occurring, this could result in schemes requiring a high level of damping, or a more severe restriction on the time-step than that determined from stability considerations only.

© 2016 Royal Meteorological Society

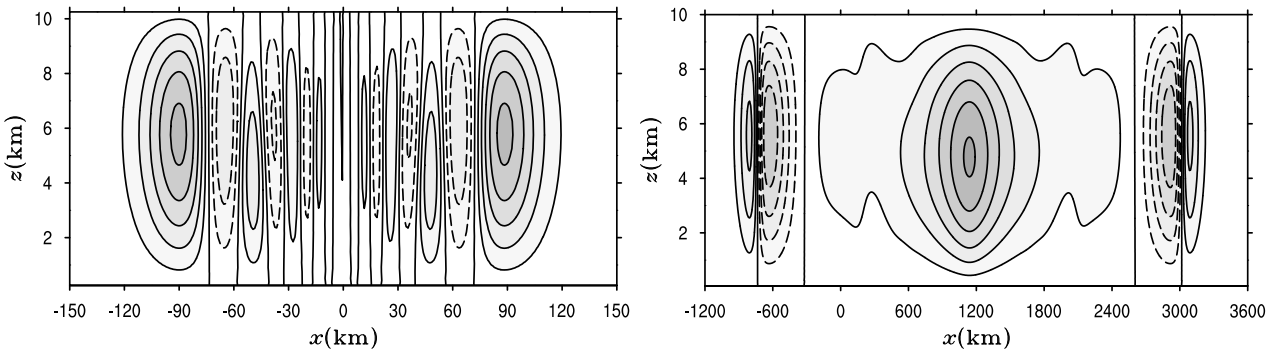


Figure 8. Mixed-Trap2(2,3,2)(-1) numerical solution of the perturbed potential temperature with $U = 20\text{m.s}^{-1}$ and $\Delta t^* = 1$, for the NHGW test-case after 3000 s (left panel), and for the HIGW test-case after 60000 s (right panel, only a part of the domain along x is represented). The contour interval is 0.5×10^{-3}K, and negative contours are dashed.

Table 1. Maximum value of basis-state Mach number $M_U = \overline{U}/\overline{c}_s(0)$ for which each scheme remains stable in NH case and H case after 48 hours of integration. Values are empirically determined to within ± 0.01; crosses indicate that instability is observed at $M_U = 0.01$.

	$\Delta t^*_{\max}$ for NHGW case:				
Scheme	1.	1.25	1.5	1.75	2.
UJ3(1,3,2)					
- UFPreF	0.71	0.42	0.16	×	×
- UFPreB	0.40	0.25	0.19	0.15	×
ARK2(2,3,2)					
- UFPreF	0.72	0.26	0.10	×	×
- UFPreB	0.93	0.70	0.31	0.28	0.25
Trap2(2,3,2)(-1)					
- UFPreF	1.00	0.65	0.37	0.15	×
- UFPreB	×	×	×	×	×
- Mixed	1.40	1.00	0.66	0.45	×
	$\Delta t^*_{\max}$ for HIGW case:				
Scheme	1.	1.25	1.5	1.75	2.
UJ3(1,3,2)					
- UFPreF	0.74	0.44	0.27	×	×
- UFPreB	0.40	0.25	0.18	0.15	×
ARK2(2,3,2)					
- UFPreF	0.79	0.27	×	×	×
- UFPreB	0.95	0.70	0.32	0.25	0.23
Trap2(2,3,2)(-1)					
- UFPreF	1.00	0.66	0.41	0.21	×
- UFPreB	×	×	×	×	×
- Mixed	1.40	1.03	0.73	0.45	0.27

In presence of advection, an even more serious drawback appears: the stability of all UFPreB variants becomes very poor; the Trap2(2,3,2)(-1) UFPreB scheme is almost unconditionally instable, being already unstable as soon as the advective wind is set to the smallest tested value 3.5 m.s^{-1}. The benefit of using UFPreB variants (advocated in LWW14) is therefore lost in presence of advection, since, for each scheme the UFPreB variant becomes less stable than its UFPreF counterpart.

The instability of UFPreB variants only when advection is present is somewhat paradoxical since the modification (from UFPreF) only involves adjustment terms. The profound reasons of this behaviour still need to be clarified, but it may be conjectured that some properties of the time-continuous evolution equations for characteristics are maintained in UFPreF variants, but violated in UFPreB. Also the fundamental mechanism leading to instability inside the core of the multi-stage algorithm of UFPreB variants is not completely elucidated here: *ad hoc* manipulations of section 5 in Butcher tableaux pointed out that the negative effect of performing the backward treatment always on the same adjustment term at each sub-stage (as in UFPreB) could be alleviated by departing from this rule, but mathematical statements ensuring stable schemes with advection are still to be found.

Drawing from all these analyses, it appears that RK-IMEX HEVI approaches based on only two Butcher tableaux are maybe not optimal for dealing with the different dynamical processes involved in a fully compressible system, and their multiple interactions.

In this prospect, a more attractive scheme based on the Trap2(2,3,2)(-1) scheme but now employing four Butcher tableaux has been proposed. In the partitioning of this mixed-Trap2(2,3,2)(-1) scheme, the two additional Butcher tableaux are specifically dedicated to the horizontal adjustment terms. Although it uses four distinct Butcher tableaux, the resulting new time-stepping scheme called mixed-Trap2(2,3,2)(-1) has been proven theoretically to be second order overall accurate, and notably, the computational cost per time-step is not larger than for the original Trap2(2,3,2)(-1) scheme. The stability domain in presence of advection is found to be more extended than for any UFPreB or UFPreF variants proposed earlier.

In this study, the Trap2(2,3,2)(-1) has been chosen to perform the extension from two to four Butcher tableaux because of the simplicity of its concept and algorithm: for instance, randomly picking up the j-th line from $\mathcal{A}^{\text{e}}$ and $\mathcal{A}^{\text{i}}$ tableaux to populate either $\mathcal{A}^{\text{u}}$ or $\mathcal{A}^{\text{p}}$ does obviously not change the second-order accuracy of the resulting scheme in such a way that the search of the most stable algorithm by this method becomes a simple optimization problem inside a limited space of permutations.

It is worth noting that similarly to this proposal, schemes with four Butcher tableaux could also be constructed based on other schemes, such as UJ3(1,3,2) or ARK2(2,3,2). However, it might be more difficult for these schemes to simultaneously satisfy the constraints linked to second-order accuracy in time, stability in presence of advection and the fulfillment of their underlying philosophy (single implicit inversion and third-order time-accuracy of the explicit part for UJ3(1,3,2); and strong stability preserving property for ARK2(2,3,2)). This has not been investigated in this paper, but the work of Kennedy and Carpenter (2003) about the construction of ARK-IMEX scheme using three Butcher tableaux for treating Convection-diffusion-reaction equations could be used as a starting basis.

From a more general point of view, this work raises questions about what should finally be the total number of Butcher tableaux

© 2016 Royal Meteorological Society

necessary to make stable (if possible) a real model with full non-linear processes (including physics), and the increased amount of stiffness brought by orographic forcing terms (as e.g. cross-terms arising with terrain-following transformed vertical coordinates commonly used in atmospheric models). The problems linked to orographic terms will be specifically examined in a forthcoming paper.

Appendix A : Double Butcher tableaux of the analysed RK-IMEX schemes

UJ3(1,3,2) :

$$\begin{array}{c|ccccc} 0 & 0 & & & & \\ 0 & 0 & 0 & & & \\ 1 & 0 & 1 & 0 & & \\ \frac{1}{2} & 0 & \frac{1}{4} & \frac{1}{4} & 0 & \\ 1 & 0 & \frac{1}{6} & \frac{1}{6} & \frac{2}{3} & 0 \\ 1 & 0 & \frac{1}{6} & \frac{1}{6} & \frac{2}{3} & 0 \\ \hline & 0 & \frac{1}{6} & \frac{1}{6} & \frac{2}{3} & 0 \end{array} \qquad \begin{array}{c|cccccc} 0 & 0 & & & & & \\ \frac{1}{2} & \frac{1}{2} & 0 & & & & \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & & & \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & & \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 & \\ 1 & \frac{1}{2} & 0 & 0 & 0 & 0 & \frac{1}{2} \\ \hline & \frac{1}{2} & 0 & 0 & 0 & 0 & \frac{1}{2} \end{array}$$

ARK2(2,3,2) :

$$\begin{array}{c|ccc} 0 & 0 & & \\ 2-\sqrt{2} & 2-\sqrt{2} & 0 & \\ 1 & \frac{1}{2}-\frac{\sqrt{2}}{3} & \frac{1}{2}+\frac{\sqrt{2}}{3} & 0 \\ \hline & \frac{\sqrt{2}}{4} & \frac{\sqrt{2}}{4} & 1-\frac{\sqrt{2}}{2} \end{array} \qquad \begin{array}{c|ccc} 0 & 0 & & \\ 2-\sqrt{2} & 1-\frac{\sqrt{2}}{2} & 1-\frac{\sqrt{2}}{2} & \\ 1 & \frac{\sqrt{2}}{4} & \frac{\sqrt{2}}{4} & 1-\frac{\sqrt{2}}{2} \\ \hline & \frac{\sqrt{2}}{4} & \frac{\sqrt{2}}{4} & 1-\frac{\sqrt{2}}{2} \end{array}$$

Trap2(2,3,2)(−1) :

$$\begin{array}{c|ccc} 0 & 0 & & \\ 1 & 1 & 0 & \\ 1 & \frac{1}{2} & \frac{1}{2} & 0 \\ 1 & \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ \hline & \frac{1}{2} & 0 & \frac{1}{2} & 0 \end{array} \qquad \begin{array}{c|ccc} 0 & 0 & & \\ 1 & 1 & 0 & \\ 1 & \frac{1}{2} & 0 & \frac{1}{2} \\ 1 & \frac{1}{2} & 0 & 0 & \frac{1}{2} \\ \hline & \frac{1}{2} & 0 & 0 & \frac{1}{2} \end{array}$$

Appendix B : Generalization of Pareschi and Russo (2005) second-order conditions for Diagonally-Implicit RK-IMEX schemes using multiples Butcher tableaux

Consider the archetype dynamical system written as a sum of N sub-systems

$$y'(t) = F[y(t),t] = \sum_{n=1}^{N} F_n[y(t),t], \tag{B.1}$$

where primes denote time-derivatives, F_n is the differential operator of the n-th sub-system, which has explicit dependencies on the state-variable $y(t)$ and, for mathematical completeness, on t itself. Through Taylor series expansion, the exact solution of (B.1) at $t+\Delta t$, is related to its value at current time t by

$$\begin{aligned} y(t+\Delta t) &= y(t) + \Delta t\, y'(t) + \frac{\Delta t^2}{2} y''(t) + O(\Delta t^3) \\ &= y(t) + \Delta t \sum_{n=1}^{N} F_n[y(t),t] + \tfrac{\Delta t^2}{2} \sum_{n=1}^{N} \partial_t F_n[y(t),t] \\ &\quad + \tfrac{\Delta t^2}{2} \sum_{n=1}^{N}\sum_{m=1}^{N} \partial_y F_n[y(t),t].F_m[y(t),t] \\ &\quad + O(\Delta t^3), \end{aligned} \tag{B.2}$$

where ∂_t is the partial time derivative, $\partial_y F_n[y(t),t]$ is the Jacobian of the multi-variable operator F_n evaluated at $[y(t),t]$, and therefore $\partial_y F_n[y(t),t].F_m[y(t),t]$ represents this Jacobian applied to $F_m[y(t),t]$.

Assume that the system (B.1) is discretized by a set of ν-stages ($\nu \geq 2$) Runge-Kutta IMEX schemes defined by N Butcher tableaux $\{\mathcal{A}^{(n)}, b^{(n)}, c^{(n)}\}$ for $n \in [1,N]$, in such a way that the Butcher tableau of index n is associated with the process F_n. Each Butcher tableau is described by the matrix $\mathcal{A}^{(n)} = \{a_{jk}^{(n)}\}$, and the weight-vectors $b^{(n)}$, $c^{(n)}$. Note that only Diagonally-Implicit RK schemes are considered herein, that is, $a_{jk}^{(n)} = 0$ for $k > j$ and $j \in [1,\nu]$. The following additional condition is assumed

$$\sum_{i=1}^{j} a_{ji}^{(n)} = c_j^{(n)}, \quad \text{for} \quad n \in [1,N]. \tag{B.3}$$

Finally, the time-discretized system writes can be written as

$$y^j = y^0 + \Delta t \sum_{n=1}^{N}\sum_{i=1}^{j} a_{ji}^{(n)} F_n(y^i, t + c_i^{(n)}\Delta t), \tag{B.4}$$

$$y^+ = y^0 + \Delta t \sum_{n=1}^{N}\sum_{j=1}^{\nu} b_j^{(n)} F_n(y^j, t + c_j^{(n)}\Delta t) \tag{B.5}$$

A multi-variable Taylor expansion of y^j using (B.4) and (B.3) leads to

$$\begin{aligned} F_n(y^j, t + c_j^{(n)}\Delta t) &= F_n(y^0,t) + c_j^{(n)}\Delta t\, \partial_t F_n(y^0,t) \\ &\quad + \Delta t \sum_{m=1}^{N} c_j^{(m)} \partial_y F_n(y^0,t).F_m(y^0,t) \\ &\quad + O(\Delta t^2) \qquad \text{for} \quad n \in [1,N]. \end{aligned} \tag{B.6}$$

Substituting (B.6) in (B.5) then yields

$$\begin{aligned} y^+ &= y^0 + \Delta t \sum_{n=1}^{N} \left(\sum_{j=1}^{\nu} b_j^{(n)} \right) F_n(y^0,t) \\ &\quad + \Delta t^2 \sum_{n=1}^{N} \left(\sum_{j=1}^{\nu} b_j^{(n)} c_j^{(n)} \right) \partial_t F_n(y^0,t) \\ &\quad + \Delta t^2 \sum_{n=1}^{N}\sum_{m=1}^{N} \left(\sum_{j=1}^{\nu} b_j^{(n)} c_j^{(m)} \right) \partial_y F_n(y^0,t).F_m(y^0,t) \\ &\quad + O(\Delta t^3) \end{aligned} \tag{B.7}$$

© 2016 Royal Meteorological Society

As a result, from (B.2) and (B.7), the scheme is second-order accurate in time provided that the coefficients of the weight-vectors of the Butcher tableaux satisfy the following conditions

$$\sum_{j=1}^{\nu} b_j^{(n)} = 1, \quad \text{for} \quad n \in [1, N], \tag{B.8}$$

$$\sum_{j=1}^{\nu} b_j^{(n)} c_j^{(m)} = \frac{1}{2}, \quad \text{for} \quad (n, m) \in [1, N] \times [1, N]. \tag{B.9}$$

These conditions can be written in a more compact form as ${}^{\mathrm{T}}b^{(n)} \cdot e = 1$ and ${}^{\mathrm{T}}b^{(n)} \cdot c^{(m)} = 1/2$ for $(n, m) \in [1, N] \times [1, N]$. Conditions (B.8) are required for first-order accuracy of the scheme, whereas conditions (B.9) for $n = m$ ensure the second-order accuracy of each Butcher tableau RK-scheme taken separately, and the cases $n \neq m$ correspond to the coupling conditions between each RK-scheme that guarantee the overall second-order accuracy of the scheme.

Appendix C : Some details of the vertical discretization and formulation of the vertically discrete Implicit problem

A Charney-Phillips vertical grid is employed where (b, w) variables are placed at grid interfaces denoted by half-integer indices $\tilde{l} \equiv l + \frac{1}{2}$, with $l \in [0, L]$; the remaining variables (u, v, P) are placed in the full-levels denoted by integer indices $l \in [1, L]$. The levels indices $1/2$ and $L + \frac{1}{2}$ refers to the bottom and the top of the domain respectively, and L denotes the total number of full levels. The vertical levels are regularly spaced in z with an increment Δz. The interpolating and differencing vertical operators are defined by

$$\overline{\langle X \rangle}^{z}_{l+\alpha} = \left(X_{l+\alpha+1/2} + X_{l+\alpha-1/2}\right)/2,$$
$$[\delta_z X]_{l+\alpha} = \left(X_{l+\alpha+1/2} - X_{l+\alpha-1/2}\right)/\Delta z,$$

where $\alpha = 1/2$ for interfaces, and $\alpha = 0$ for full-levels. Thus, we define by $D^z_{\tilde{l}} P_l = [\delta_z P]_{l+1/2} + \overline{\langle \Gamma_w P \rangle}^{z}_{l+1/2}$ the vertical operator acting on P in the w-vertical momentum equation, and by $D^z_l w_{\tilde{l}} = [\delta_z w]_l - \left[(1-\kappa)/\overline{H}_l\right] \overline{\langle w \rangle}^{z}_{l}$ the vertical operator acting on w in pressure equation. Rigid material conditions are imposed at top and bottom boundaries assumed flat, so that $w_{1/2} = w_{L+1/2} = 0$. Besides, a free-slip boundary condition is assumed for the horizontal wind, hence $u_{1/2} = u_1$ and $u_{L+1/2} = u_L$.

The implicit problem arising at each implicit sub-stage is solved by deriving a vertically discrete Helmholtz equation for the single variable w, as detailed below. At each implicit sub-stage $j \in [1, \nu]$, the vertically-discrete implicit problem writes

$$u_l^j + \Delta t a^{\mathrm{p}}_{jj} \nabla P_l^j = u_l^{j\bullet} \tag{C.1}$$

$$w_{\tilde{l}}^j - \Delta t a^{\mathrm{i}}_{jj} b_l^j + \Delta t a^{\mathrm{i}}_{jj} D^z_{\tilde{l}} P_l^j = w_{\tilde{l}}^{j\bullet} \tag{C.3}$$

$$b_l^j + \Delta t a^{\mathrm{i}}_{jj} \overline{N}^2 w_{\tilde{l}}^j = b_l^{j\bullet} \tag{C.4}$$

$$P_l^j + \Delta t a^{\mathrm{i}}_{jj} \overline{c}^2_{s\,l} D^z_l w_{\tilde{l}}^j + \Delta t a^{\mathrm{u}}_{jj} \overline{c}^2_{s,l} \nabla u_l^j = P_l^{j\bullet} \tag{C.5}$$

where ∇ is the spectrally computed horizontal derivative, and the terms $(u_l^{j\bullet}, w_{\tilde{l}}^{j\bullet}, b_l^{j\bullet}, P_l^{j\bullet})$ are known quantities which include contributions from the explicit part of the j-th sub-stage. The formalism here uses the coefficients $(a^{\mathrm{e}}, a^{\mathrm{i}}, a^{\mathrm{u}}, a^{\mathrm{p}})$ of the proposed mixed scheme with four Butcher tableaux, being understood that classical UFPreF, UFPreB and UBPreF schemes can also be represented by this formalism, simply replacing them by $(a^{\mathrm{e}}, a^{\mathrm{i}}, a^{\mathrm{e}}, a^{\mathrm{e}})$, $(a^{\mathrm{e}}, a^{\mathrm{i}}, a^{\mathrm{i}}, a^{\mathrm{e}})$ and $(a^{\mathrm{e}}, a^{\mathrm{i}}, a^{\mathrm{e}}, a^{\mathrm{i}})$ respectively. Since the Coriolis terms are treated explicitly, v equation is not part of the implicit problem.

Now, proceeding by successive eliminations in favour of $w_{\tilde{l}}^j$, and using the fact that under HEVI condition $a^{\mathrm{p}}_{jj} a^{\mathrm{u}}_{jj} = 0$, leads to a single tridiagonal vertical problem for w^j of the form

$$w_{\tilde{l}}^j - \frac{\Delta t^2 a^2_{jj}}{1 + \Delta t^2 a^2_{jj} \overline{N}^2} D^z_{\tilde{l}} \left(\overline{c}^2_{s\,l} D^z_l \, w_{\tilde{l}}^j \right) = \frac{w_{\tilde{l}}^{j\bullet} + \Delta t a_{jj} b_{\tilde{l}}^{j\bullet}}{1 + \Delta t^2 a^2_{jj} \overline{N}^2} - \frac{\Delta t a_{jj}}{1 + \Delta t^2 a^2_{jj} \overline{N}^2} D^z_{\tilde{l}} \left(P_l^{j\bullet} - \Delta t a^{\mathrm{u}}_{jj} \overline{c}^2_{s\,l} \nabla u_l^{j\bullet} \right) \tag{C.6}$$

This equation, supplemented with $w_{1/2} = w_{L+1/2} = 0$ can be solved using a tridiagonal matrix algorithm. Once w^j is determined, b^j is obtained from (C.4) and the determination of P^j and u^j is proceeds as follows

- In UFPreB variant, u^j is first computed explicitly from (C.1), and P^j is determined by back-substitution of w^j and u^j in (C.5).
- In UBPreF variant, P^j is first computed explicitly from (C.5) and u^j is updated by back-substitution of P^j in (C.1).
- In the case of the Mixed-Trap2(2,3,2) scheme, the procedure is alternatively that of UBPreF scheme and that of UFPreB scheme.
- In UFPreF variant, u^j and P^j are directly from (C.1) and (C.5).

Acknowledgements

The authors thanks Dr. Hilary Weller and another anonymous reviewer for their contructive comments which helped to improve the manuscript.

References

Ascher, U. M., S. J. Ruuth, and R. J. Spiteri, 1997: Implicit-explicit Runge-Kutta methods for time-dependent partial differential equations. *Appl. Numer. Math.*, **25**, 151–167,

Durran D.R., Blossey P.N., 2012. Implicit-Explicit multistep methods for fast-wave-slow-wave problems. *Mon. Weather Rev.* **140 (4)**: 1307–1325.

Giraldo F., and Restelli M., 2008: A study of spectral element and discontinuous Galerkin methods for the Navier-Stokes equations in nonhydrostatic mesoscale atmospheric modeling: Equation sets and test cases. *J. Comput. Phys.*, **227**, 3849–3877

Giraldo F., J.F. Kelly, and E. Constantinescu. 2013. Implicit-Explicit formulations of a three-dimensional nonhydrostatic unified model of the atmosphere (NUMA). *SIAM J. Comput.* **35 (5)**: 1162–1194.

Gottlieb S., C.-W. Shu, and E. Tadmor, 2001. Strong stability-preserving high-order time discretization methods. *SIAM rev.* **43**: 89–112.

Kennedy C., and Carpenter M., 2003. Additive Runge-Kutta schemes for convection-diffusion-reaction equations, *Appl. Numer. Math.* **44**: 139–181.

Klemp J.B., Wilhelmson R.B., 1978. The simulation of three-dimensional convective storm dynamics. *J. Atmos. Sci.* **35 (6)**: 1070–1096.

Lock S.-J., N. Wood, and Weller H., 2014. Numerical analyses of Runge-Kutta implicit-explicit schemes for horizontally explicit, vertically implicit solutions of atmospheric models. *Q. J. R. Meteorol. Soc.* **140 (682)**: 1654–1669.

Melvin T., Dubal M., Wood N., Staniforth A., Zerroukat M., 2010. An inherently mass-conserving iterative semi-implicit semi-Lagrangian discretization of the non-hydrostatic vertical-slice equations. *Q. J. R. Meteorol. Soc.* **136**: 799–814.

Mesinger F., 1977. Forward-Backward scheme, and its use in a limited area model. *Contributions to Atmospheric Physics* **50**: 200–210.

Pareschi L., Russo G., 2001. Implicit-explicit Runge-Kutta schemes for stiff systems of differential equations. *Adv. Theory Comput Math.* **3**: 269–289.

© 2016 Royal Meteorological Society

Pareschi L., Russo G., 2005. Implicit-Explicit Runge-Kutta schemes and applications to hyperbolic systems with relaxation. *J. Sci. comput.* **25 (1-2)**: 129–155.

Satoh M., 2002. Conservative scheme for the compressible nonhydrostatic models with the Horizontally Explicit and Vertically Implicit time integration scheme. *Mon. Weather Rev.* **130 (5)**: 1227–1245.

Skamarock W.C., Klemp J.B., 1994. Efficiency and accuracy of the Klemp-Wilhelmson time-splitting technique. *Mon. Weather Rev.* **122 (11)**: 2623–2630.

Ullrich P., and Jablonowski C., 2012. Operator-split Runge-Kutta-Rosenbrock methods for nonhydrostatic atmospheric models. *Mon. Weather. Rev.* **140 (4)**: 1257–1284.

Weller H., S.-L. Lock, and Wood N., 2013. Runge-Kutta IMEX schemes for the Horizontally Explicit/Vertically Implicit (HEVI) solution of wave equations. *J. Comput. Phys.* **252**: 365–381.

Wicker LJ, Skamarock WC. 2002. Time splitting methods for elastic models using forward time schemes. *Mon. Weather Rev.* **130**: 2088–2097.

Wood N., Staniforth A., White A., Allen T., Diamantakis M., Gross M., Melvin T., Smith C., Vosper S., Zerroukat M., and Thuburn J., 2013. An inherently mass-conserving semi-implicit semi-Lagrangian discretisation of the deep-atmosphere global non-hydrostatic equations. *Q. J. R. Meteorol. Soc.* **252**: 365–381.

© 2016 Royal Meteorological Society

Annexe B

Algorithme de discrimination des phases Acoustiques-Gravité

Pour discriminer parmi les quatre valeurs propres $\lambda_l\, l \in [\![1;4]\!]$, définies par les schémas d'intégration temporels, lesquelles font référence aux ondes acoustiques (notées λ_a) ou de gravité (notées λ_g), nous allons utiliser une propriété toujours vérifiée pour les schémas que nous utilisons à savoir que, sauf cas exceptionnels, ces quatre valeurs s'accouplent en deux paires de nombres complexes conjugués. Cette propriété montre donc bien que les ondes numériques se propagent à la même vitesse, dans la même direction, mais dans un sens opposé. Ainsi, il est facile d'isoler les deux paires. Pour cela, il suffit de faire toutes les sommes possibles $\mathrm{Im}(\lambda_i + \lambda_j)$ et de créer la paire $\{\lambda_i; \lambda_j\}$ si cette somme est inférieure à $\epsilon = 10^{-10}$. Reste à savoir quelle est la paire modélisant des ondes acoustiques, et celle des ondes de gravité. Pour cela, il suffit de faire une évaluation des phases : la paire ayant la plus grande (en valeur absolue) est celle faisant référence aux ondes acoustiques, et l'autre celle qui renvoie aux ondes de gravité.

Il faut noter que dans le cas des ondes acoustiques, cette opposition entre les phases n'est pas toujours respectée.Dans ce cas précis, nous faisons jouer l'argument de continuité sur les valeurs propres en fonction de C_*. De là, nous observons que la paire qui reste complexe prolonge dans la continuité la valeur des ondes de gravité. Nous concluons donc que la paire réelle correspond aux ondes acoustiques. il est très difficile de différencier l'onde positive de l'onde négative. Dans ce cas, un choix arbitraire de notre part sera effectué. Cet algorithme est résumé par le schéma suivant :

Données : $\{\lambda_1, \lambda_2, \lambda_3, \lambda_4\}$
Résultat : $\{\lambda_a, \lambda_g\}$
pour tous les $(i, j) \in [\![1, 4]\!] \times [\![1, 4]\!]$ **faire**
  si $(i \neq j)$ **et** $(|\mathrm{Im}(\lambda_i) + \mathrm{Im}(\lambda_j)| < \epsilon)$ **alors**
    $\mathtt{tmp}_1 \leftarrow (\lambda_i, \lambda_j)$;
    pour tous les $k \in [\![1, 4]\!]$ **faire**
      si $(k \neq i)$ **et** $(k \neq j)$ **alors**
        $\mathtt{tmp}_2 \leftarrow (\lambda_k, \lambda_{10-i-j-k})$;
      fin
    fin
  fin
fin
$\mathtt{Normal} \leftarrow$ **Vrai** ;
si $|\mathrm{Im}(\mathtt{tmp}_1[1]) + \mathrm{Im}(\mathtt{tmp}_1[2])| < \epsilon)$ **alors**
  $\lambda_a \leftarrow \mathtt{tmp}_1$;
  $\mathtt{Normal} \leftarrow$ **Faux** ;
  si $\mathrm{Im}(\mathtt{tmp}_2[1] > \mathrm{Im}(\mathtt{tmp}_2[2])$ **alors**
    $\lambda_g \leftarrow (\mathtt{tmp}_2[1], \mathtt{tmp}_2[2])$;
  sinon
    $\lambda_g \leftarrow (\mathtt{tmp}_2[2], \mathtt{tmp}_2[1])$;
  fin
fin
si $|\mathrm{Im}(\mathtt{tmp}_2[1]) + \mathrm{Im}(\mathtt{tmp}_2[2])| < \epsilon)$ **alors**
  $\lambda_a \leftarrow \mathtt{tmp}_2$;
  $\mathtt{Normal} \leftarrow$ **Faux** ;
  si $\mathrm{Im}(\mathtt{tmp}_1[1] > \mathrm{Im}(\mathtt{tmp}_1[2])$ **alors**
    $\lambda_g \leftarrow (\mathtt{tmp}_1[1], \mathtt{tmp}_1[2])$;
  sinon
    $\lambda_g \leftarrow (\mathtt{tmp}_1[2], \mathtt{tmp}_1[1])$;
  fin
fin
si $\mathtt{Normal}$ **alors**
  si $\max\{\mathrm{Arg}(\mathtt{tmp}_1)\} > \max\{\mathrm{Arg}(\mathtt{tmp}_2)\}$ **alors**
    si $\mathrm{Im}(\mathtt{tmp}_1[1] > \mathrm{Im}(\mathtt{tmp}_1[2])$ **alors**
      $\lambda_c \leftarrow (\mathtt{tmp}_1[1], \mathtt{tmp}_1[2])$;
    sinon
      $\lambda_c \leftarrow (\mathtt{tmp}_1[2], \mathtt{tmp}_1[1])$;
    fin
    si $\mathrm{Im}(\mathtt{tmp}_2[1] > \mathrm{Im}(\mathtt{tmp}_2[2])$ **alors**
      $\lambda_g \leftarrow (\mathtt{tmp}_2[1], \mathtt{tmp}_2[2])$;
    sinon
      $\lambda_g \leftarrow (\mathtt{tmp}_2[2], \mathtt{tmp}_2[1])$;
    fin
  sinon
    si $\mathrm{Im}(\mathtt{tmp}_2[1] > \mathrm{Im}(\mathtt{tmp}_2[2])$ **alors**
      $\lambda_c \leftarrow (\mathtt{tmp}_2[1], \mathtt{tmp}_2[2])$;
    sinon
      $\lambda_c \leftarrow (\mathtt{tmp}_2[2], \mathtt{tmp}_2[1])$;
    fin
    si $\mathrm{Im}(\mathtt{tmp}_1[1] > \mathrm{Im}(\mathtt{tmp}_1[2])$ **alors**
      $\lambda_g \leftarrow (\mathtt{tmp}_1[1], \mathtt{tmp}_1[2])$;
    sinon
      $\lambda_g \leftarrow (\mathtt{tmp}_1[2], \mathtt{tmp}_1[1])$;
    fin
  fin
fin

Annexe C

Algorithmes d'inversion de matrices à bandes

Algorithme d'inversion d'une matrice tridiagonale

Soit le problème matriciel tridiagonal suivant :

$$\begin{pmatrix} c_1 & d_1 & 0 & & & 0 & \\ b_2 & \ddots & \ddots & \ddots & & & \\ 0 & \ddots & & & & & \\ & \ddots & & & & & \\ & & & & & & 0 \\ & & & & & & d_{L-1} \\ 0 & & & & 0 & b_L & c_L \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ \\ \\ \\ x_L \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ \\ \\ \\ y_L \end{pmatrix}$$

avec c_j non-nul pour tout j compris entre 1 et L.

Pour résoudre cette équation, Llewellyn H. Thomas propose une version simplifiée du pivot de Gauss. L'idée est d'effectuer une première descente pour éliminer les coefficients sous-diagonaux b_j en les exprimant en fonction des autres coefficients de la matrice. Puis, lors de la remontée, toutes les inconnues sont exprimées par des relations de récurrence. Concrètement, l'algorithme s'écrit de la manière suivante :

Données : $\{a, b, c, d, e, y\}$
Résultat : x
$d'_1 \leftarrow d_1/c_1$;
$y'_1 \leftarrow y_1/c_1$;
pour tous les $j \in [\![2; L-1]\!]$ **faire**
  $\texttt{tmp} \leftarrow c_j - b_j d'_{j-1}$;
  $d'_j \leftarrow d_j/\texttt{tmp}$;
  $y'_j \leftarrow (y_j - b_j y'_{j-1})/\texttt{tmp}$;
fin
$x_L = y'_L$;
pour tous les $j \in [\![1; L-2]\!]$ **faire**
  $x_{L-1-j} \leftarrow (y'_{L-1-j} - d'_{L-1-j} * x_{L-j}$;
fin

Alors que de manière générale, le pivot de Gauss est un algorithme d'ordre $O(L^3)$, cet algorithme n'est que d'ordre $O(L)$. Cette économie de calcul nécessite néanmoins la condition que les éléments diagonaux de la matrice soient non-nuls. Dans le cas considéré ici, ces éléments sont strictement positifs, ce qui assure la faisabilité de cet algorithme.

Algorithme d'inversion d'une matrice pentadiagonale

Considérons maintenant un problème matriciel s'écrivant de la manière suivante :

$$\begin{pmatrix} c_1 & d_1 & e_1 & 0 & & 0 & \\ b_2 & \ddots & \ddots & \ddots & & & \\ a_3 & \ddots & & & & & \\ 0 & \ddots & & & & & 0 \\ & & & & & & e_{L-2} \\ & & & & & & d_{L-1} \\ 0 & & & 0 & a_L & b_L & c_L \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ \\ \\ \\ x_L \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ \\ \\ \\ y_L \end{pmatrix}$$

avec c_j non-nul pour tout j compris entre 1 et L.

Pour résoudre ce type d'équation, et en suivant l'idée de l'algorithme de double descente, il suffit de modifier les coefficients de la matrice par un premier passage pour écrire l'ensemble des inconnues par une relation de récurrence. Puis, lors du second passage, en déduire l'ensemble des variables.

Données : $\{a, b, c, d, e, y\}$
Résultat : x
pour tous les $j \in [\![2; L-1]\!]$ **faire**
  $\texttt{tmp} \leftarrow b_j / c_{j-1}$;
  $c_j \leftarrow c_j - \texttt{tmp} * d_{j-1}$;
  $d_j \leftarrow d_j - \texttt{tmp} * e_{j-1}$;
  $y_j \leftarrow y_j - \texttt{tmp} * y_{j-1}$;
  $\texttt{tmp} \leftarrow a_{j+1} / c_{j-1}$;
  $b_{j+1} \leftarrow b_j - \texttt{tmp} * d_{j-1}$;
  $c_{j+1} \leftarrow c_j - \texttt{tmp} * e_{j-1}$;
  $y_{j+1} \leftarrow y_{j+1} - \texttt{tmp} * y_{j-1}$;
fin
$\texttt{tmp} \leftarrow b_L / c_{L-1}$;
$c_L \leftarrow c_L - \texttt{tmp} * d_{L-1})$;
$x_L \leftarrow (y_L - \texttt{tmp} * y_{L-1}) / c_L$;
$x_{L-1} \leftarrow (y_{L-1} - d_{L-1} * x_L) / c_{L-1}$;
pour tous les $j \in [\![1; L-2]\!]$ **faire**
  $x_{L-1-j} \leftarrow (y_{L-1-j} - e_{L-1-j} * x_{L+1-j} - d_{L-1-j} * x_{L-j}) / c_{L-1-j}$;
fin

Comparé à la méthode de double descente classique, cet algorithme est également d'ordre $O(L)$, mais nécessite plus de calculs. Elle permet une résolution directe du problème, sans avoir à réaliser de décomposition de type LU.

Bibliographie

[1] Akio Arakawa et Celal S. Konor : Unification of the anelastic and quasi-hydrostatic systems of equations. *Monthly Weather Review*, 137(2):710–726, 2009.

[2] Uri M. Ascher, Steven J. Ruuth et Raymond J. Spiteri : Implicit-Explicit Runge-Kutta methods for time-dependent partial differential equations. *Applied Numerical Mathematics*, 25(2):151–167, 1997.

[3] Uri M. Ascher, Steven J. Ruuth et Brian T.R. Wetton : Implicit-Explicit methods for time-dependent partial differential equations. *SIAM Journal on Numerical Analysis*, 32:797–823, 1995.

[4] Richard Asselin : Frequency filter for time integrations. *Monthly Weather Review*, 100:487–490, 1972.

[5] Piere Bénard, Jozef Vivoda, Ján Mašek, Petra Smolíková, Karim Yessad, Ch Smith, Radmila Brožková et Jean-François Geleyn : Dynamical kernel of the Aladin–NH spectral limited-area model : Revised formulation and sensitivity experiments. *Quarterly Journal of the Royal Meteorological Society*, 136(646):155–169, 2010.

[6] Pierre Bénard : Stability of Semi-Implicit and iterative centered-implicit time discretizations for various equation systems used in NWP. *arXiv preprint physics/0311123*, 2002.

[7] Pierre Bénard, René Laprise, Jozef Vivoda et Petra Smolíková : Stability of Leap-Frog constant-coefficients semi-implicit schemes for the fully elastic system of Euler equations : Flat-terrain case. *arXiv preprint physics/0311123*, 2003.

[8] Pierre Bénard, Ján Mašek et Petra Smolíková : Stability of Leap-Frog constant-coefficients semi-implicit schemes for the fully elastic system of Euler equations : Case with orography. *Monthly Weather Review*, 133(5):1065–1075, 2005.

[9] Radmila Bubnová, Gwenaëlle Hello, Pierre Bénard et Jean-François Geleyn : Integration of the fully elastic equations cast in the hydrostatic pressure terrain-following coordinate in the framework of the ARPEGE/Aladin NWP system. *Monthly Weather Review*, 123(2):515–535, 1995.

[10] Jules G Charney, Ragnar Fjörtoft et John Von Neumann : Numerical integration of the barotropic vorticity equation. *Tellus*, 2(4):237–254, 1950.

[11] James W. Cooley et John W. Tukey : An algorithm for the machine calculation of complex Fourier series. *Mathematics of computation*, 19(90):297–301, 1965.

[12] Richard Courant, Kurt Friedrichs et Hans Lewy : Über die partiellen Differenzengleichungen der mathematischen Physik. *Mathematische annalen*, 100(1):32–74, 1928.

[13] MMichael J.P. Cullen : Alternative implementations of the semi-Lagrangian semi-implicit schemes in the ECMWF model. *Quarterly Journal of the Royal Meteorological Society*, 127(578):2787–2802, 2001.

[14] Roger Daley : The normal modes of the spherical non-hydrostatic equations with applications to the filtering of acoustic modes. *Tellus A*, 40(2):96–106, 1988.

[15] Huw C. Davies : Limitations of some common lateral boundary schemes used in regional NWP models. *Monthly Weather Review*, 111(5):1002–1012, 1983.

[16] Thomas Dubos et Marine Tort : Equations of atmospheric motion in Non-Eulerian vertical coordinates : Vector-Invariant form and Quasi-Hamiltonian formulation. *Monthly Weather Review*, 142(10):3860–3880, 2014.

[17] Thomas Dubos et Fabrice Voitus : A semihydrostatic theory of gravity-dominated compressible flow. *Journal of the Atmospheric Sciences*, 71(12):4621–4638, 2014.

[18] Dale R. Durran : Improving the anelastic approximation. *Journal of the atmospheric sciences*, 46(11):1453–1461, 1989.

[19] Dale R. Durran : The third-order Adams-Bashforth method : an attractive alternative to Leap-Frog time differencing. *Monthly Weather Review*, 119(3):702–720, 1991.

[20] Dale R. Durran et Peter N. Blossey : Implicit-Explicit multistep methods for fast-wave-slow-wave problems. *Monthly Weather Review*, 140(4):1307–1325, 2012.

[21] John A. Dutton : The Ceaseless Wind, 579 pp, 1986.

[22] Arnt Eliassen : *The quasi-static equations of motion with pressure as independent variable*, volume 17. Grøndahl & Sons boktr., I kommisjon hos Cammermeyers boghandel, 1949.

[23] Tzvi Gal Chen et Richard C.J. Somerville : On the use of a coordinate transformation for the solution of the Navier-Stokes equations. *Journal of Computational Physics*, 17(2):209–228, 1975.

[24] Almut Gassmann et Hans-Joachim Herzog : A consistent time-split numerical scheme applied to the nonhydrostatic compressible equations. *Monthly Weather Review*, 135(1):20–36, 2007.

[25] Francis X. Giraldo, John F. Kelly et Emil Constantinescu : Implicit-Explicit formulations of a three-dimensional nonhydrostatic unfied model of the atmosphere (NUMA). *SIAM Journal on Computing*, 35(5):1162–1194, 2013.

[26] Francis X. Giraldo et Marco Restelli : A study of spectral element and discontinuous galerkin methods for the Navier–Stokes equations in nonhydrostatic mesoscale atmospheric modeling : Equation sets and test cases. *Journal of Computational Physics*, 227(8):3849–3877, 2008.

[27] Sigal Gottlieb et Chi-Wang Shu : Total variation diminishing Runge-Kutta schemes. *Mathematics of Computation of the American Mathematical Society*, 67(221):73–85, 1998.

[28] Sigal Gottlieb, Chi-Wang Shu et Eitan Tadmor : Strong stability-preserving high-order time discretization methods. *SIAM review*, 43(1):89–112, 2001.

[29] Anthony Hollingsworth : *A spurious mode in the "Lorenz" arrangement of Ø and T which does not exist in the "Charney-Phillips" arrangement*. European Centre for Medium-Range Weather Forecasts, 1995.

[30] Motohki Ikawa : Comparison of some schemes for nonhydrostatic models with orography. *Journal Meteorological Society of Japan*, 66:753–776, 1988.

[31] Meriem Jedouaa, Charles-Henri Bruneau et Emmanuel Maitre : An efficient interface capturing method for a large collection of interacting cells immersed in a fluid. 2015.

[32] Akira Kasahara : Various vertical coordinate systems used for numerical weather prediction. *Monthly Weather Review*, 102(7):509–522, 1974.

[33] Christopher A. Kennedy et Mark H. Carpenter : Additive Runge-Kutta schemes for convection-diffusion-reaction equations. Rapport technique, NASA Langley Technical Report Server, 2001.

[34] Joseph B. Klemp, William C. Skamarock et Jimy Dudhia : Conservative split-explicit time integration methods for the compressible nonhydrostatic equations. *Monthly Weather Review*, 135(8):2897–2913, 2007.

[35] Joseph B. Klemp et Robert B. Wilhelmson : The simulation of three-dimensional convective storm dynamics. *Journal of the Atmospheric Sciences*, 35(6):1070–1096, 1978.

[36] Jean-Philippe Lafore, Joël Stein, Nicole Asencio, Philippe Bougeault, Véronique Ducrocq, Jacqueline Duron, Claude Fischer, Philippe Héreil, Patrick Mascart, Valéry Masson *et al.* : The Meso-NH atmospheric simulation system. Part I : Adiabatic formulation and control simulations. *In Annales Geophysicae*, volume 16, pages 90–109. Springer, 1997.

[37] Lev Davidovitch Landau et Evgenii Mikhailovich Lifchitz : *Physique théorique : Mécanique des fluides*, volume 6. Éditions Mir, ellipses-edition marketing édition, 1971.

[38] Lev Davidovitch Landau et Evgenii Mikhailovich Lifchitz : *Physique théorique : Mécanique*, volume 1. Éditions Mir, Ellipses-Edition Marketing édition, 1994.

[39] René Laprise : The Euler equations of motion with hydrostatic pressure as an independent variable. *Monthly Weather Review*, 120(1):197–207, 1992.

[40] Randall J. Leveque et Joseph Oliger : Numerical methods based on additive splittings for hyperbolic partial differential equations. *Mathematics of computation*, 40(162):469–497, 1983.

[41] Franik B. Lipps et Richard S. Hemler : A scale analysis of deep moist convection and some related numerical calculations. *Journal of the Atmospheric Sciences*, 39(10):2192–2210, 1982.

[42] Sarah-Jane Lock, Nigel Wood et Hilary Weller : Numerical analyses of Runge–Kutta implicit–explicit schemes for horizontally explicit, vertically implicit solutions of atmospheric models. *Quarterly Journal of the Royal Meteorological Society*, 140(682):1654–1669, 2014.

[43] Ytzhaq Mahrer : An improved numerical approximation of the horizontal gradients in a terrain-following coordinate system. *Monthly Weather Review*, 112(5):918–922, 1984.

[44] Glenn Mainguy : *L'économie du quotidien. Une étude de la précarité à travers l'exemple des pratiques agricoles domestiques dans le monde rural russe*. Thèse de doctorat, Université de Bordeaux, 2016.

[45] Gurii I. Marchuk : Numerical solution of problems of atmosphere and ocean dynamics. *Gidrometeoizdat, Leningrad*, 1974.

[46] Thomas Melvin, Mark Dubal, Nigel Wood, Andrew Staniforth et Mohamed Zerroukat : An inherently mass-conserving iterative semi-implicit semi-Lagrangian discretization of the non-hydrostatic vertical-slice equations. *Quarterly Journal of the Royal Meteorological Society*, 136(648):799–814, 2010.

[47] Fedor Mesinger : Forward–Backward scheme, and its use in a limited area model. *Contrib. Atmos. Phys.*, 50:200–210, 1977.

[48] Eike H. Müller et Robert Scheichl : Massively parallel solvers for elliptic partial differential equations in numerical weather and climate prediction. *Quarterly Journal of the Royal Meteorological Society*, 140(685):2608–2624, 2014.

[49] Yoshimitsu Ogura et Norman A. Phillips : Scale analysis of deep and shallow convection in the atmosphere. *Journal of the atmospheric sciences*, 19(2):173–179, 1962.

[50] Lorenzo Pareschi et Giovanni Russo : Implicit-Explicit Runge-Kutta schemes and applications to hyperbolic systems with relaxation. *Journal of Scientific computing*, 25(1-2):129–155, 2005.

[51] Alexandre Philip, Thierry Bergot, Yves Bouteloup et François Bouyssel : The impact of vertical resolution on fog forecasting in the kilometric-scale model AROME : A case study and statistics. *Weather and Forecasting*, (2016), 2016.

[52] Norman A. Phillips : A coordinate system having some special advantages for numerical forecasting. *Journal of Meteorology*, 14(2):184–185, 1957.

[53] R. James Purser et Lance M. Leslie : Reducing the error in a time-split finite-difference scheme using an incremental technique. *Monthly Weather Review*, 119(2):578–585, 1991.

[54] André J. Robert : The integration of a low order spectral form of the primitive meteorological equations (Spherical harmonics integration of low order spectral form of primitive meteorological equations). *Journal Meteorological Society of Japan*, 44:237–245, 1966.

[55] André J. Robert, John Henderson et Colin Turnbull : An implicit time integration scheme for baroclinic models of the atmosphere. *Monthly Weather Review*, 100(5):329–335, 1972.

[56] Masaki Satoh : Conservative scheme for the compressible nonhydrostatic models with the Horizontally Explicit and Vertically Implicit time integration scheme. *Monthly Weather Review*, 130(5):1227–1245, 2002.

[57] Christoph Schär, Daniel Leuenberger, Oliver Fuhrer, Daniel Lüthi et Claude Girard : A new terrain-following vertical coordinate formulation for atmospheric prediction models. *Monthly Weather Review*, 130(10):2459–2480, 2002.

[58] Juan Simarro et Mariano Hortal : A semi-implicit non-hydrostatic dynamical kernel using finite elements in the vertical discretization. *Quarterly Journal of the Royal Meteorological Society*, 138(664):826–839, 2012.

[59] Adrian J. Simmons, Brian J. Hoskins et David M. Burridge : Stability of the semi-implicit method of time integration. *Monthly Weather Review*, 106(3):405–412, 1978.

[60] William C. Skamarock et Joseph B. Klemp : The stability of time-split numerical methods for the hydrostatic and the nonhydrostatic elastic equations. *Monthly Weather Review*, 120(9): 2109–2127, 1992.

[61] William C. Skamarock et Joseph B. Klemp : Efficiency and accuracy of the Klemp-Wilhelmson time-splitting technique. *Monthly Weather Review*, 122(11):2623–2630, 1994.

[62] Ronald B Smith : Linear theory of stratified hydrostatic flow past an isolated mountain. *Tellus*, 32(4):348–364, 1980.

[63] Piotr K. Smolarkiewicz, Christian Kühnlein et Nils P. Wedi : A consistent framework for discrete integrations of soundproof and compressible pdes of atmospheric dynamics. *Journal of Computational Physics*, 263:185–205, 2014.

[64] Jerry M. Straka, Robert B. Wilhelmson, Louis J. Wicker, John R. Anderson et Kelvin K. Droegemeier : Numerical solutions of a non-linear density current : A benchmark solution and comparisons. *International Journal for Numerical Methods in Fluids*, 17(1):1–22, 1993.

[65] Gilbert Strang : On the construction and comparison of difference schemes. *SIAM Journal on Numerical Analysis*, 5(3):506–517, 1968.

[66] Yasuo Tatsumi : An economical explicit time integration scheme for a primitive model (of atmosphere). *Journal Meteorological Society of Japan*, 61:269–288, 1983.

[67] Paul Ullrich et Christiane Jablonowski : Operator-split Runge-Kutta-Rosenbrock methods for nonhydrostatic atmospheric models. *Monthly Weather Review*, 140(4):1257–1284, 2012.

[68] Hilary Weller, Sarah-Jane Lock et Nigel Wood : Runge-Kutta IMEX schemes for the Horizontally Explicit/Vertically Implicit (HEVI) solution of wave equations. *Journal of Computational Physics*, 2013.

[69] Louis J. Wicker et William C. Skamarock : A time-splitting scheme for the elastic equations incorporating second-order Runge-Kutta time differencing. *Monthly Weather Review*, 126(7):1992–1999, 1998.

[70] Louis J. Wicker et William C. Skamarock : Time-splitting methods for elastic models using forward time schemes. *Monthly Weather Review*, 130(8):2088–2097, 2002.

[71] Paul D. Williams : A proposed modification to the Robert-Asselin time filter. *Monthly Weather Review*, 137(8):2538–2546, 2009.

[72] Paul D. Williams : The RAW filter : an improvement to the Robert-Asselin filter in semi-implicit integrations. *Monthly Weather Review*, 139(6):1996–2007, 2011.

[73] Nigel Wood, Andrew Staniforth, Andy White, Thomas Allen, Michail Diamantakis, Markus Gross, Thomas Melvin, Chris Smith, Simon Vosper, Mohamed Zerroukat *et al.* : An inherently mass-conserving semi-implicit semi-Lagrangian discretization of the deep-atmosphere global non-hydrostatic equations. *Quarterly Journal of the Royal Meteorological Society*, 140(682):1505–1520, 2014.

[74] Günther Zängl : Extending the numerical stability limit of terrain-following coordinate models over steep slopes. *Monthly Weather Review*, 140(11):3722–3733, 2012.